# உலகத் தமிழ் இணையக் கருத்தரங்க மாநாடு

# TamilNet 99

7, 8 பிப்ரவரி 1999

கட்டுரைத் தொகுப்பு

*Conference Papers*

தகவல் தொழில்நுட்பத்துறை
தமிழ்நாடு அரசு, சென்னை,
இந்தியா.

கட்டுரைத் தொகுப்பு

(CONFERENCE PAPERS)

# உலகத் தமிழ் இணையக் கருத்தரங்க மாநாடு

# TamilNet 99

Organised by
Information Technology Department
Government of Tamil Nadu
Chennai, INDIA

7 & 8 February 1999

*Allied Publishers Limited*
Ahmedabad • Mumbai • Bangalore • Calcutta • Hyderabad
Lucknow • Chennai • Nagpur • New Delhi

© February 1999
Electronics Corporation of Tamil Nadu Ltd.
Chennai - 600 025, India

**Allied Publishers Limited**
Prarthna Flats (1st Floor), Navrangpura, Ahmedabad - 380 009
15, J.N. Heredia Marg, Ballard Estate, Mumbai - 400 038
3-5-1114, Kachiguda Cross Road, Hyderabad - 500 027
81, Hill Road, Ramnagar, Nagpur - 440 010
Patiala House, 16A Ashok Marg, Lucknow - 226 001
5th Main Road, Gandhi Nagar, Bangalore - 560 009
17, Chittaranjan Avenue, Calcutta - 700 072
13/14, Asaf Ali Road, New Delhi - 110 002
751, Mount Road, Chennai - 600 002

ISBN 81-7023-903-6

*Typeset at*
VISVALAYA
9, Velders Street, Anna Salai
Chennai - 600 002
✆ 8570778

*Printed at*
HITECH OFFSET (P) LIMITED,
No. 1, Angamuthu Street,
Royapettah,
Chennai 600 014
✆ 8546681

*Published by*
Mr. Sunil Sachedev
M/s. ALLIED PUBLISHERS LTD.

## MEMBERS OF STATE LEVEL TASK FORCE FOR INFORMATION TECHNOLOGY INDUSTRY IN TAMIL NADU AS PER G.O. MS. NO. 652(Industries(MIE-2)Dept. dt. 08.10.1998.

| | |
|---|---|
| **HON'BLE CHIEF MINISTER** | **CHAIRMAN** |
| Thiru Murasoli Maran M.P. | Vice- Chairman |
| Chief Secretary | Member |
| Secretary, Finance Department | Member |
| Secretary, Industries Department | Member-Convenor |
| Secretary, Information Technology Department | Member |
| Dr. M. Anandakrishnan<br>(Former Vice-Chancellor, Anna University)<br>Vice-Chairman,<br>Tamil Nadu State Council for Higher Education. | Member |
| Chairman and Managing Director ELCOT | Member |
| Thiru Dewang Mehta<br>(Executive Director, NASSCOM) | Member |
| Thiru Ram N. Agarwal<br>President,<br>Manufacturers' Association for IT | Member |
| Thiru K.V. Ramani,<br>Managing Director.<br>Future Software Pvt. Ltd. | Member |
| Thiru S. Mahalingam<br>Executive President,<br>Tata Consultancy Services | Member |
| Thiru Gopal Srinivasan<br>Director,<br>TVS Electronics | Member |
| Dr. C.R. Muthukrishnan<br>Prof. of Computer Science,<br>Indian Institute of Technology | Member |

## MEMBERS OF THE SUB-COMMITTEE FOR INTRODUCTION OF TAMIL SOFTWARE AND INCLUSION OF INFORMATION TECHNOLOGY IN THE SCHOOL CURRICULUM AS PER G.0.MS. NO. 653, Industries(MIE-2) Department Dt. 08.10.1998.

| | |
|---|---|
| Dr. M. Anandakrishnan<br>(Former Vice-Chancellor, Anna University)<br>Vice-Chairman,<br>Tamil Nadu State Council for Higher Education. | Chairman |
| Secretary, Information Technology | Member-Convenor |
| Secretary, School Education | Member |
| Dr. M. Ponnavaikko<br>Professor, Computer Science & Engg.<br>Creseent Engineering College, Chennai | Member |
| Dr. K. Sarukesi<br>Professor, Computer Science,<br>Bharathiyar University, Coimbatore. | Member |
| Thiru. K. M. Mohanasundaram<br>Professor, Computer Science,<br>PSG College of Technology,<br>Coimbatore. | Member |
| Dr. V.Krishnamurthy<br>Professor,<br>School of Computer Science & Engineering,<br>Anna University, Chennai. | Member |
| Dr. P. Tamizh Chelvi<br>Reader,<br>Iyya Nadar Janaki Ammal College,<br>Chinna Kammanpatti, Sivakasi-Est. | Member |
| Dr. C. Xavier<br>Lecturer in Computer Science,<br>St. Xavier College, Palayankottai | Member |
| Thiru S. Rangarajan(Sujatha)<br>Writer, Chennai | Member |

## RECEPTION COMMITTEE MEMBERS FOR TamilNet 99
## As per G.O. Ms. No.3 Information Technology (M12) Dept. dt. 2.12.1998.

| | |
|---|---|
| Thiru Murasoli Maran M.P.<br>(Vice-Chairman, State Level IT<br>Task Force) | Chairman |
| Hon'ble Minister for Tamil OL&TC &HR CE | Member |
| Hon'ble Minister for Information and<br>Publicity | Member |
| Hon'ble Minister for Tourism | Member |
| Secretary, Industries Department | Member |
| Secretary, Public Department | Member |
| Dr. M. Anandakrishnan<br>(Former Vice-Chancellor, Anna University)<br>Vice-Chairman,<br>Tamil Nadu State Council for Higher Education. | Member |
| Thiru S. Rangarajan (Sujatha)<br>Writer, Chennai | Member |
| Secretary, Information Technology | Member-Secretary |

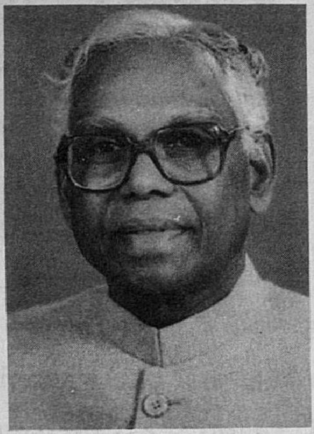
**Secretariat -TamilNet 99**

Thiru K.S. Lakshminarayanan
General Manager(Development)
Chief Technical Advisor
ELCOT

सत्यमेव जयते

राष्ट्रपति
भारत गणतंत्र

PRESIDENT
REPUBLIC OF INDIA



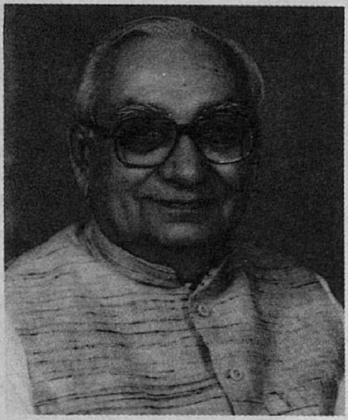
## MESSAGE

The International Conference and Seminar on Tamil in Information Technology - "TamilNet 99" - at Chennai, will play a useful role in promoting the use of Tamil on computers in many parts of the world.

I extend my warm greetings to the distinguished participants and organisers and wish the deliberations every success.

(K.R. NARAYANAN)

February 3, 1999
New Delhi



सत्यमेव जयते

भारत के उप-राष्ट्रपति

VICE-PRESIDENT OF INDIA

# MESSAGE

I am happy to learn that the Government of Tamil Nadu is organising an International Conference and Seminar on Tamil in Information Technology, captioned `TamilNet 99', at Chennai on 7-8 February, 1999.

The revolution in information technology in this century has posed a major challenge to the oriental languages as well as presented them with an unprecedented opportunity. These languages, structured according to precise norms of grammar and syntax, are capable of being adapted to computer with the minimum effort. Tamil, as one of India's most ancient, mature and rich languages, can surely lead the way in computer adaptability. This will enable people conversant in Tamil language to freely use computers and benefit from the ever escalating information technology.

I am glad that the Government of Tamil Nadu has taken a very thoughtful step in the systematic evolution of Tamil for computer usage aimed at development of internationally acceptable standards.

I wish the effort every success.

Krishan Kant

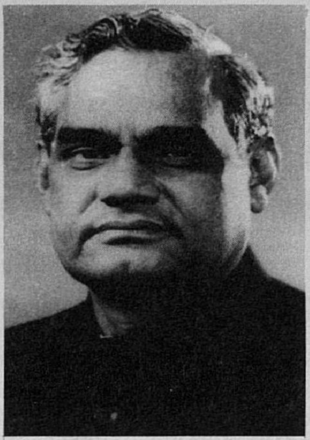
(KRISHAN KANT)

New Delhi.
29th January, 1999.

सत्यमेव जयते

प्रधान मंत्री

**PRIME MINISTER**



## MESSAGE

I am pleased to know that an international conference and seminar on Tamil in Information Technology, "TamilNet99", is being organised in Chennai on February 7-8, 1999. I have said that IT is India's Tomorrow. However to ensure that all Indians are able to use its many benefits, there is an urgent need to increase local language content in information technology. Tamil is a classical language, whose richness is enhanced manifold by its ability to adapt itself to the challenges of the modern era. Tamil Nadu has produced a large number of highly successful Information Technology professionals and entrepreneurs, who are today operating both in India and abroad.

It is very heartening to note that the Information Technology Department of the Government of Tamil Nadu has organised this seminar to take Information Technology for the common people by enabling Tamil to become a major player in the Information revolution.

I send my best wishes to the seminar and all its participants.

ABVajpayee

(A.B. Vajpayee)

New Delhi
February 2, 1999

**மு. கருணாநிதி**
முதலமைச்சர்

தலைமைச் செயலகம்
சென்னை - 600 009

கணிப்பொறி தொழில் நுட்பமும், **'இண்டர்நெட்'** என்று கூறப்படுகிற இணையத் தொடர்புகளும் பெருமளவுக்கு வளர்ந்து வரும் இன்றைய நிலையில்; இந்த வளர்ச்சிகளில் தமிழ் மொழியை முழு அளவில் ஈடுபடுத்த வேண்டிய தேவை உருவாகியுள்ளது. இதன் அடிப்படையில் வரும் 7 மற்றும் 8 தேதிகளில் சென்னையில் **'தமிழ்நெட் 99'** என்கிற **உலகத் தமிழ் இணையக் கருத்தரங்க மாநாடு** நடத்த தமிழ்நாடு அரசு முடிவு எடுத்துள்ளது.

பல நாடுகளிலிருந்து வரும் வல்லுநர்களின் கருத்துக்களைக் கண்டறிந்து உலகமெங்கும் கணிப்பொறிகளிலும் இணையத்திலும் தமிழ் மொழியைச் சீராகப் பயன்படுத்தவும்; அதற்கேற்ப நவீன மென் பொருள்களை உருவாக்கவும்; தமிழ்கூறும் நல்லுலகில் வாழும் அனைவர்க்கும் பொதுவான தரப்படுத்தப்பட்ட அடிப்படை குறியீட்டு முறைகளை நிர்ணயிக்கவும்; மாநாடு வழி செய்யும். அதன் மூலம் இண்டர்நெட் 'தமிழ்நெட்டாக' மாறும் நிலை உருவாகும்.

அப்போது தான் சாதாரண மனிதனுக்கும், கிராமத்தில் வாழ்வோருக்கும் கணிப்பொறி ஒரு அர்த்தமுள்ள தோழனாக – தொண்டனாக – கல்வி கற்பிக்கும் ஆசானாகப் பயன்பட முடியும்.

அதற்காக இதுவரை மேற்கொள்ளப்பட்ட முயற்சிகளையும், இனி ஏற்பட இருக்கும் முயற்சிகளையும் ஒருங்கிணைக்கும் ஒரு மாபெரும் பணியைத் தான் இம்மாநாடு துவங்கி வைக்கிறது.

வரலாறு படைக்கும் இப்பணியினைத் தொடர்ந்து, தமிழ்நாடு அரசு அடுத்த நூற்றாண்டில் தகவல் தொழில் நுட்பத் துறையில் புரட்சிகரமான முன்னேற்றங்களுக்கும், தமிழ் அறிந்த வல்லுநர்கள் செயல்பட எல்லாவித ஊக்கங்களுக்கும் வழி வகுக்கும்.

வெளிநாட்டிலிருந்து வரும் உயர் குழு வல்லுநர்களுடன் தமிழ் நாட்டில் உள்ள வல்லுநர்களும் கலந்தாலோசித்துத் தொடர்பு கொள்ளவும், அத்தகைய தொடர்பினை நீட்டிக்கவும், பொதுவான முடிவுகளை மேற்கொள்ளவும் என் வாழ்த்துக்களைக் குவிக்கிறேன்.

அன்புள்ள,

(மு. கருணாநிதி)

# முன்னுரை

உலகெங்கும் பரவி வாழ்ந்து வரும் தமிழர்களிடையே நிலையான தொடர்புகள் வளர இன்றைய தகவல் தொழில் நுட்பம் வழி காட்டுகின்றது. தமிழ் நாட்டின் பொருளாதார வளர்ச்சியிலும், சமூக முன்னேற்றத்திலும், ஆர்வம் கொண்ட பல்லாயிரக்கணக்கான தமிழர்கள் பல நாடுகளில் உள்ளனர். தகவல் தொழில்நுட்பத்தின் மூலம் தமிழ் இலக்கியத்தை அறியவும், தமிழ்க் கல்வியை வளர்க்கவும் பலர் முயற்சிசெய்து வருகின்றனர். இவற்றுக்கெல்லாம் பேருதவி தர வல்லது இணையத்தின் வளர்ச்சியாகும்.

ஆர்வத்தின் அடிப்படையில் தனித்தனியாக இம்முயற்சிகளில் ஈடுபடும்போது, இணையத்தின் முழு பயனையும் தமிழ் உலகம் பெற முடியாமல் போகுமோ என்ற ஐயம் எழுந்துள்ளது. தமிழ் நாட்டில் கணிப்பொறியை பெரும் அளவிற்கு பயன்படுத்த வேண்டிய தேவைகள் நாளுக்கு நாள் பெருகி வரும் நிலையில், கணிப்பொறியில் தமிழைப் பயன்படுத்த வேண்டிய கட்டாய நிலை உருவாகி உள்ளது. இதற்காகத் தேவைப்படும் தமிழ் மென்பொருள்களும், விசைப்பலகைகளும் அவரவர் விருப்பப்படி உருவாக்குவதன் மூலம் தகவல் பரிமாற்றங்களில் பல வித சிக்கல்கள் எழுகின்றன.

கணிப்பொறியிலும், இணையத்திலும் இன்றைய நிலையில் தமிழின் செயல்பாடுகளையும், வளர்ச்சிகளையும் கண்டறியவும், எதிர்காலத்தில் மேலும் பல முன்னேற்றங்களுக்கு வழிவகுக்கவும் உலகத்தமிழ் இணைபக் கருத்தரங்க மாநாடு வாய்ப்பளிக்கின்றது.

இந்தக் கருத்தரங்கின் முக்கிய மூன்று குறிக்கோள்கள்: (1) தமிழ் விசைப்பலகையின் சீரமைப்பு (2) தமிழ் எழுத்துக்களுக்கான கணிப்பொறி குறியீடுகளின் ஒருமைப்பாடு. (3) வருங்கால தமிழ் இணைய வளர்ச்சி. இந்த மூன்று பொருள்களைச் சார்ந்த ஆய்வுக்கட்டுரைகளை இந்தத் தொகுப்பு நூல் மூலம் மகிழ்ச்சியுடன் அளிக்கின்றோம். இம்மாநாட்டில் கருதப்படவிருக்கும் அடிப்படைப் பொருட்களில் எல்லா அம்சங்களையும் தீர ஆலோசித்து இதை செவ்வனே நடத்துவதற்கு முழு உற்சாகத்தையும் பேருதவியையும் அளித்தவர் தமிழக முதலமைச்சர் கலைஞர் கருணாநிதி அவர்கள். இம்மாநாட்டின் வரவேற்புக் குழுவிற்குத் தலைமை தாங்கி இதன் நிகழ்ச்சிகள் சிறப்பாக அமைய வேண்டி எல்லா நடவடிக்கைகளையும் அயராது இரவு பகலாக மேற்பார்வை செய்து அறிவுரைகள் வழங்கியவர் பாராளுமன்ற உறுப்பினர் திரு. முரசொலி மாறன் அவர்கள்.

குறுகிய காலத்தில் பொருள் செறிந்த கட்டுரைகளை அளித்த அறிஞர்களுக்கு மிக்க நன்றி. இந்த ஆய்வுக்கட்டுரைகளை சீர் தூக்கிப் பார்த்து, ஒழுங்குபடுத்தி அச்சிட உதவி செய்த முனைவர் பொன்னவைக்கோ, முனைவர் கிருஷ்ணமூர்த்தி, திரு. லட்சுமிநாராயணன் கொண்ட குழுவினருக்கு எமது பாராட்டுக்கள். இந்தத் தொகுப்பை சிறிய கால அளவில் அச்சிட்டு வெளியிட உதவிய அலைடு பதிப்பகத்திற்கு நன்றி. இம்முயற்சிகளுக்கு அயராது உதவி அளித்த தகவல் தொழில் நுட்பத்துறை செயலர் தா. பிரகாஷ் அவர்கள் பெரும் பாராட்டுக்குரியவர்.

**மு. ஆனந்தகிருஷ்ணன்**

*துணைத் தலைவர்,*

*தமிழ்நாடு மாநில உயர்கல்வி மன்றம்.*

# CONTENTS

## SECTION - I
*Standardisation of Tamil Key-Board*


1.1. தமிழ்க் கணிப்பொறி விசைப்பலகை இடஅமைவுகளின் ஓர் ஒப்பாய்வு — 3
*மு. பொன்னவைக்கோ, சென்னை*

1.2. Computer Keyboard for Tamil — 12
*V. Krishnamorthy*, Chennai

1.3. Tamil in Cyberspace — 18
*Ramalingam Shanmugalingam*, USA


## SECTION - II
*Standardisation of Coding of Tamil Scripts*


2.1. Towards a Total Internet Solutions for the Tamil Language through Singapore Research — 25
*Naa. Govindaswamy*, Singapore

2.2. A Glyph based Font Encoding Scheme TSCII as a Candidate for Tamil Computing — 32
*K. Kalyanasundaram*, Switzerland &
*M. Nedumaran*, Singapore

2.3. Unified Coding Scheme for Tamil — 38
*V. Krishnamorthy*, Chennai

2.4. Multiple Encoding System for Different Computer Applications — 41
*M. Ganesan*, Mysore

2.5. A Proposal for Tamil Standard Code for Information interchange — 46
*N. Anbarasan*, Banagalore

2.6. Tamil, Tamizhee, and Nadodi — 51
*Manoj Annadurai*, Chennai &
*Benjamin Arun Kumar*

2.7. Tamil Font Encoding Standards – A Critical Study — 58
*P. Chellappan*, Chennai

2.8. Proposal for a Standard Encoding of Tamil Characters — 64
*Sinnathurai Srivas*, USA

2.9. iDNS, a DNS System with Mutilingual Support — 70
*James Seng.*, Singapore

2.10. Computing with Tamil : Use of the vernacular in Designing user interfaces — 71
*Kalyanakrishnan*, Chennai

2.11. Towards Defining the Design Goals while placing Tamil in Uni-code — 79
*P. Chandrabose*, Chennai

2.12. தமிழ் நெடுங்கணக்கில் புதியன புகுதல் 82
செ.ச. செந்தில்நாதன், சென்னை


## SECTION - III
*Use of Tamil in Multimedia Internet and Computing*

3.1. A Tamil Speech Synthesis Sytem 89
*A. G. Ramakrishnan*, Banagalore

3.2. உலகளாவிய மின் வலையில் தமிழ்ப் பாடம் 95
அரங்க வாசு, பிலடெல்பியா, (USA)

3.3. Project Madurai for E-texts of Tamil Literary Works 102
*K. Kalyanasundarm*, Switzerland

3.4. Computer - Aided Identification of Tamil Words 108
*V.M. Andavar*, Chennai & *V. M. Kavi Arasan*,

3.5. கணிமொழி – 'சி' 112
மு. பொன்னவைக்கோ, சென்னை

3.6. Senthamil Learning and Teaching Tamil Through Multimedia 118
*Anbarasan*, Bangalore

3.7. தமிழ் ஆராய்ச்சிக்குக் கைகொடுக்கும்
தகவல் தொழில்நுட்பம் 122
சு. சீனிவாசன், கல்பாக்கம்

3.8. தொன்ம வேர்களை இணைக்கும் மின் வேர்! 127
சா.அ. சௌரிராசன்

*SECTION - I*

# Standardisation of Tamil Key-Board

# தமிழ்க் கணிப்பொறி விசைப்பலகை இடஅமைவுகளின் ஓர் ஒப்பாய்வு

**Dr.மு.பொன்னவைக்கோ**
**தலைமைக் கணிப்பொறி பேராசிரியர்**
**கிரசன்ட் பொறியியற் கல்லூரி**
**சென்னை – 600048**

## முன்னுரை

கணிப்பொறியின் வளர்ச்சியும் பயன்பாடும் நாளுக்கு நாள் பெருகிக் கொண்டே வருகின்றது. கணிப்பொறி எல்லாத் துறையினருக்கும் ஒரு இன்றியமையாத கருவியாக வளர்ந்துள்ளதோடு மனிதனின் அன்றாட வாழ்க்கைக்கும் தேவையான சாதனமாக வளர்ந்து வருகின்றது. நகர மக்களோடு நின்றுவிடாமல் கிராம மக்களின் பயன்பாட்டிலும் கணிப்பொறி இடம் பெற்று வருகின்றது. கணிப்பொறியை பயன்படுத்த மிக முக்கியமான சாதனம் உள்ளீட்டு கருவியாகிய விசைப்பலகையாகும். கணிப்பொறிக்கு உள்ளீடு செய்வோர் பயிற்சி பெற்ற தட்டச்சாளர்கள் மட்டுமின்றி பல்வகைப்பட்ட தட்டச்சுப் பயிற்சி பெறாத கணிப்பொறி பயனாளர்களுமாவர். பயிற்சி பெற்றவர்களை விட பயிற்சி பெறாதவர்களே எண்ணிக்கையில் மிகுதியாவர். எனவே கணிப்பொறி விசைப்பலகையில் எழுத்துக்களின் இட அமைவு இயக்குவதற்கு எளிமையாகவும், இலக்கணமரபு கொண்டதாகவும், திறம்மிக்கதாகவும், விரலசைவுகள் குறைந்த நிலையிலும் மிகவிரைவில் நிறைந்த தகவல்களை உள்ளீடு செய்யவல்லதாகவும் இருத்தல் வேண்டும்.

இன்று கணிப்பொறியில் தமிழின் பயன்பாடு உலக அளவில் மிகப்பேரளவில் வளர்ந்துள்ளது. சிங்கப்பூர், மலேசியா, இலங்கை, அமெரிக்கா, கனடா மற்றும் பல நாடுகளிலும் கணிப்பொறித்தமிழ், தனிக்கணிப்பொறி முதல் இணையம் வரை வளர்ச்சி பெற்றுள்ளது. உலகின் பல பகுதிகளிலுமுள்ள ஆய்வாளர்கள் பல வகைப்பட்ட கணிப்பொறித் தமிழ் விசைப்பலகைகளை வடிவமைத்து அளித்துள்ளனர். பல்வேறுபட்ட இட அமைவுகளைக் கொண்ட விசைப்பலகைகள் பயன்பாட்டில் உள்ளன. இதனால் ஒன்றுக்கொன்று தொடர்பு இன்றி கணிப்பொறி பயனாளர்களுக்கு பெரும் சிக்கலை உருவாக்கியுள்ளது. இந்த சிக்கலிலிருந்து விடுபட கணிப்பொறி விசைப்பலகையின் இட அமைவுகளை தரப்படுத்த வேண்டிய நிலை உருவாகியுள்ளது. இந்த முயற்சிக்கு உதவும் வகையில், இக்கட்டுரையில், இதுவரை வடிவமைக்கப்பட்டுள்ள விசைப்பலகைகளின் ஓர் ஒப்பாய்வு விவரிக்கப்படுகின்றது.

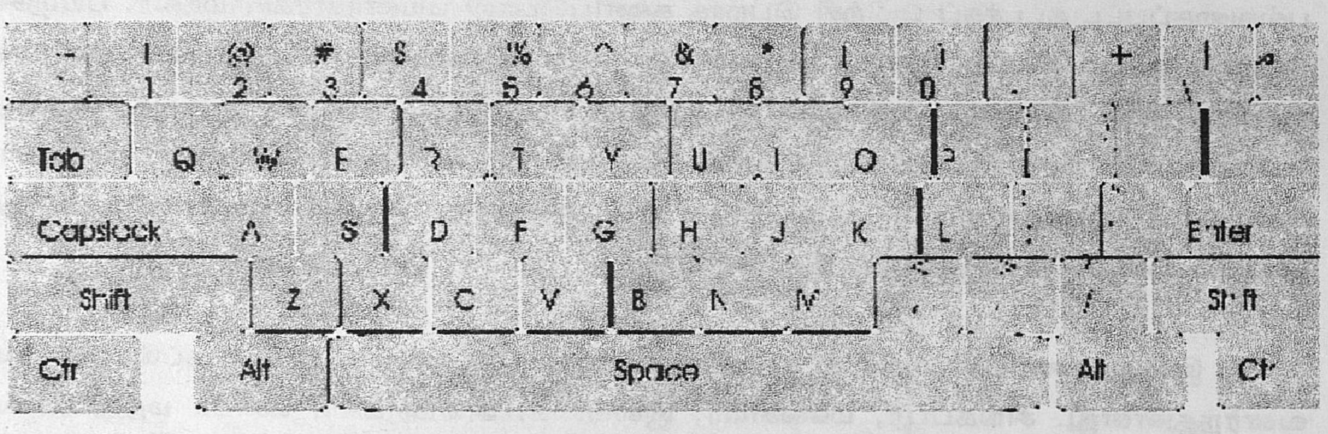
### விசைப்பலகையின் வகைகள்

இதுவரை வடிவமைக்ககப்பட்டுள்ள கணிப்பொறி விசைப்பலகைகளை மூன்று வகைகளாகப் பிரிக்கலாம். அவையாவன,

1. ரோமன் எழுத்து விசைப்பலகைகள்
2. தட்டச்சு வகை விசைப்பலகைகள்
3. ஒலியியல் வகை விசைப்பலகைகள்

### ரோமன் எழுத்து விசைப்பலகை

முதன்மதலில் கணிப்பொறியில் தமிழை பயன்படுத்த முயன்ற அனைவரும் கையாண்ட உள்ளீட்டு முறை ரோமன் எழுத்து முறையாகும். இவ்வகையில், எழுத்துப் பெயர்ப்பு முறையில் தமிழுக்கு இணையான ரோமன் எழுத்துக்களைத் தெரிவு செய்து பயன்படுத்தப்படுகிறது. இவ்வகை விசைப்பலகையை படம் - 1 - இல் காணலாம்




படம் - 1 ரோமன் எழுத்து விசைப்பலகை

இவ்வகை விசைப்பலகைளை தமிழ் பல்கலைக்கழகம் திருக்குறள் பகுப்பாய்விலும் கனடாவில் உள்ள திரு.K.சீனிவாசன் தனது 'ஆதவின்' என்னும் சொல் தொகுப்பியிலும், மலேசியாவில் உள்ள முத்தெழிலன் தனது 'முரசு' என்னும் சொல் தொகுப்பியிலும் பயன்படுத்தியுள்ளனர்.

ரோமன் எழுத்துவகை விசைப்பலகையை இயக்க நிறைந்த முயற்சிகள் தேவை. இதன் இடஅமைவுகள் இயல்பானவை அல்ல; இதனை இயக்க ஆங்கில அறிவு தேவை.

### தட்டச்சு வகை விசைப்பலகைகள்

இரண்டு தட்டச்சு வகை விசைப்பலகைகள் உருவாக்கப்பட்டுள்ளன. ஒன்று 'மரபு வகை'; மற்றது சுவிட்சர்லாந்தில் உள்ள திரு.K.கல்யாண சுந்தரம் உருவாக்கிய 'மயிலை'. மரபு வகை படம் - 2 - லும், 'மயிலை' படம் - 3 - லும் காண்க. மரபுவகை விசைப்பலகையில் கொம்பு, கால், அகர மேறிய உயிர்மெய் எழுத்துக்கள் போன்ற ஒவ்வொன்றுக்கும் ஒரு சாவியை இயக்க வேண்டும். ஒகரம், ஓகாரம் போன்ற எழுத்துக்களை உள்ளிட மூன்று சாவிகளை தனித்தனியே இயக்க வேண்டும். 'மயிலை' விசைப்பலகையில் 'Shift' சாவியைப் பயன்படுத்தி உகர, சில ஊகார எழுத்துக்களை உள்ளிட வகை செய்யப்பட்டுள்ளது. இவ்வகையிலும் கொம்பு, கால் போன்ற துணை எழுத்துக்களுக்குத் தனித்தனி சாவிகள் உள்ளன.

தட்டச்சு வகை விசைப்பலகைகள் ஆங்கிலம், தமிழ் ஆகிய இரு மொழிகளிலும் தட்டச்சுப் பயிற்சிபெற்றோர்க்கு எளிதாக இருக்கும். ஆயின் இவ்வகையைப் பயன்படுத்த மிகுந்த முயற்சியும் பயிற்சியும் தேவை. தட்டச்சுப் பயிற்சியில்லா கணிப்பொறிபயனாளர்கட்டு எளிமையானதல்ல.

### ஒலியியல் வகை விசைப்பலகைகள்

இந்திய மொழிகளின் எழுத்து அமைப்புகள் தனிச்சிறப்பு வாய்ந்தவை. உயிர் எழுத்துக்களும் மெய் எழுத்துக்களும் மொழியின் அடிப்படையாகும். இவை புணர்வதால் உயிர்மெய் எழுத்துக்களாகிய கூட்டு எழுத்துக்கள் உருவாகின்றன. இந்த அடிப்படைத் தத்துவத்தைப் பயன்படுத்தி வடிவமைக்கப்பட்டதே ஒலியியல் வகை விசைப்பலகைகளாகும். இந்திய மொழிகளிலேயே தமிழ்மொழி ஒரு தனித்தன்மை பெற்றதாகும்.

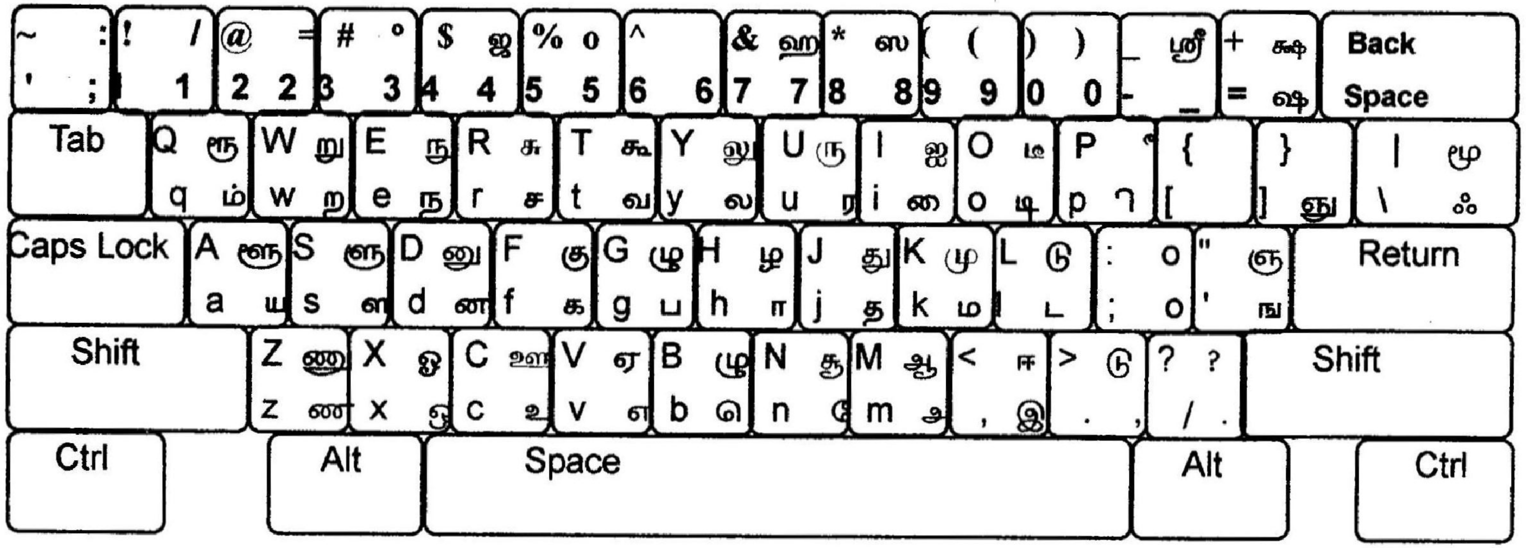
பல ஒலியியல் வகை விசைப்பலகைகள் வடிவமைக்கப்பட்டுள்ளன. இந்திய மொழிகளின் பயன்பாட்டுக்கென CDAC (Centre for Development of Advanced Computing) எனப்படும் மத்திய அரசு நிறுவனம் 'INSCRIPT' என்னும் விசைப்பலகையை வடிவமைத்துள்ளது (படம் - 4). சிங்கப்பூரைச் சேர்ந்த Dr.நா.கோவிந்தசாமி 'கணியன்' என்ற விசைப்பலகையை வடிவமைத்து பயன்படுத்திவருகிறார் (படம் - 5). இலங்கைத் தமிழர்கள் சிங்கள, ஆங்கில மொழிகளுடன் பயன்படக்கூடிய ஒரு தமிழ் விசைப்பலகையை வடிவமைத்துள்ளனர் (படம் - 6). தினமலர் திரு.R.கிருஷ்ணமூர்த்தி ஒரு விசைப்பலகையை வடிவமைத்து வழங்கியுள்ளார் (படம் - 7). அண்ணா பல்கலைக் கழகத்தைச் சேர்ந்த கணிப்பொறித்துறை பேராசிரியர் Dr.வெ.கிருஷ்ணமூர்த்தி 'அறந்தை' என்ற பெயரில்

ஒரு விசைப்பலகையை வடிவமைத்து வழங்கியுள்ளார் (படம் - 8). தமிழக அரசால் அமைக்கப்பட்ட விசைப்பலகை தரக்குழு மூன்று வகையான விசைப்பலகைகளை வடிவமைத்து பரிந்துரை செய்துள்ளது (படம் - 9, 10, 11) இவற்றில் கடைசி இரண்டைத் தவிர பிற அனைத்தும் ஒலியியல் ஒமைப்பில் வடிவமைக்கப்பட்ட விசைப்பலகைகளாகும்.

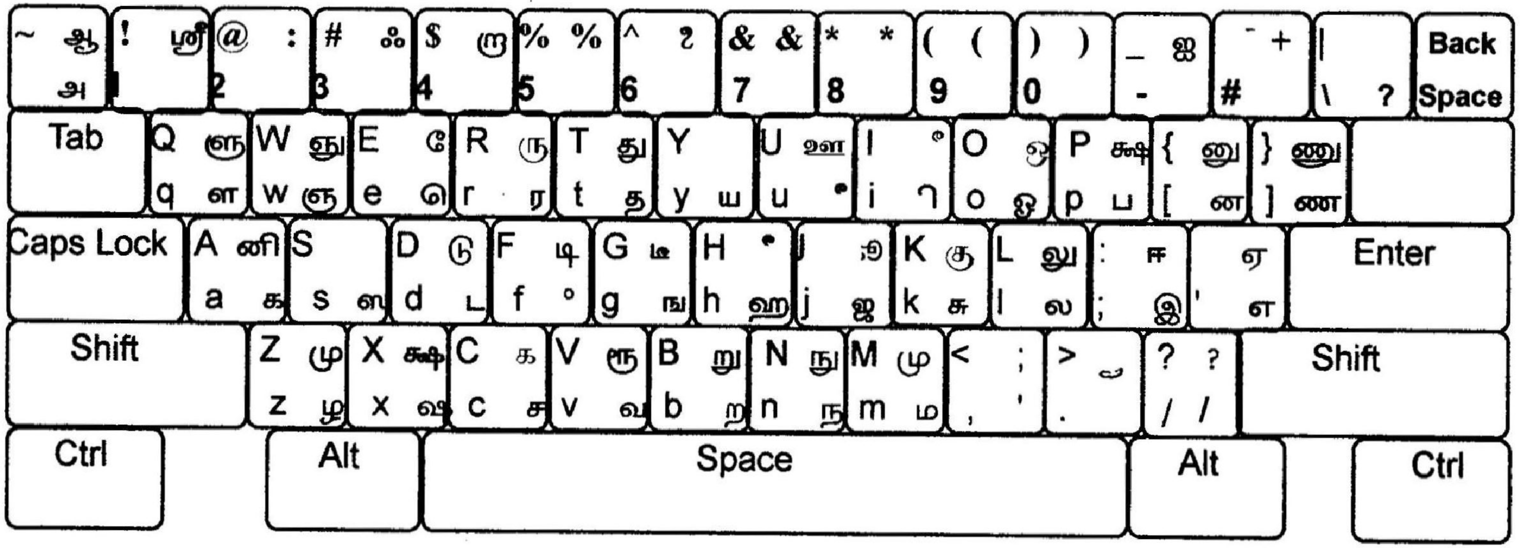
மத்திய அரசு நிறுவனமாகிய 'C DAC' அமைத்துள்ள 'Inscript' விசைப்பலகை, அனைத்து இந்திய மொழிகளுக்கும் பயன்படுமாறு ஒரு பொதுவான விசைப்பலகையாக 1980 இல் வடிவமைக்கப்பட்டு பயன்பாட்டில் இருந்துவருகிறது. தமிழ் மொழி பிற இந்திய மொழிகளிலிருந்து வேறுபட்ட தனிச்சிறப்புடைய மொழியாகையால், பொதுவான Inscript விசைப்பலகையில் தமிழின் சிறப்பு அம்சங்கள் இடம்பெறவில்லை.

'கணியன்' விசைப்பலகை எழுத்து புழக்கத்தை அடிப்படையாக கொண்டு அமைக்கப்பட்டது. இவ்வமைப்பில், தமிழில் மிகுதியாக பயன்படுத்தக்கூடிய எழுத்துக்களாகிய க, த, அ போன்ற எழுத்துக்களை ஆள்காட்டி விரல் கொண்டு தட்டும்வகையிலும் குறைவாகப் பயன்படக்கூடிய எழுத்துக்களுக்கு அவற்றின் பயன்பாட்டிற்கு ஏற்ப பிற விரல்கள் இயங்கும் இடங்களிலும் இட ஒதுக்கீடு செய்யப்பட்டுள்ளன. இதனால் இவ்வகை விசைப்பலகையைக் கொண்டு பிற விசைப்பலகைகளைவிட, ஒரு குறிப்பிட்ட நேரத்தில், மிகுதியான தகவல்களை உள்ளீடு செய்ய முடியும்.

அண்ணா பல்கலைக்கழகத்தை சேர்ந்த Dr.வெ.கிருஷ்ணமூர்த்தி வடிவமைத்து வழங்கியுள்ள அறந்தை விசைப்பலகை பல சிறப்பு அம்கங்களை பெற்றுள்ளது. இதில் விரல்களுக்கு சுமை ஏறக்குறைய சமமாக பகிர்ந்து அளிக்கப்பட்டுள்ளது. தமிழில் இன எழுத்துக்களாகிய ந்த, ண்ட, ன்ற, ம்ப போன்ற எழுத்துக்களை அடையாளம் கண்டு வல்லின எழுத்துக்களுக்கு முன் வரும் இன எழுத்துக்களின் மெய் எழுத்துக்களான புள்ளி தானே இடப்படுகிறது. இதனால், ஒரு இயக்கம் மிச்சப்படுகிறது. மேலும் வல்லின எழுத்துக்களாகிய க ச ட த ப ற, இருமுறை ஒரு சேர வந்தால் முன் வரும் எழுத்து மெய் எழுத்தாக புள்ளியை தானே பெருகின்றது. எடுத்துக்காட்டாக 'கக' என அச்சிட்டால் 'க்க' என தானே அச்சிடப்படுகிறது. இதனால் குறைந்த முயற்சியில் நிறைந்த தகவல்களை மிக எளிதாக உள்ளிட முடிகின்றது. இதுவரை வடிவமைக்கப் பட்டுள்ள விசைப்பலகைகளை விட இது பல சிறப்பு அம்சங்களைப் பெற்றுள்ளது. இந்த விசைப்பலகையில் உயிர் எழுத்துக்களுக்கான சாவிகள் வலப்பக்கமும் அகர மேறிய உயிர் மெய் எழுத்துக்களுக்கான சாவிகள் இடப்பக்கமாகவும் அமைக்கப் பட்டுள்ளன. தட்டச்சு வல்லுனர்களின் கருத்து இதற்கு மாறுபட்டதாக உள்ளது. உயிர் எழுத்துக்கள் இடப்புறமும் அகர மேறிய மெய் எழுத்துக்கள் வலப்புறமும் அமைவது சிறப்பானதென கருதப்பபடுகிறது. அறந்தை விசைப்பலகையில் இந்த மாற்றம் செயல் படுத்தப்பட்டால் இது மிகச்சிறந்த விசைப்பலகையாக அமையும்.




படம் - 2 தட்டச்சு வகை (மரபு)

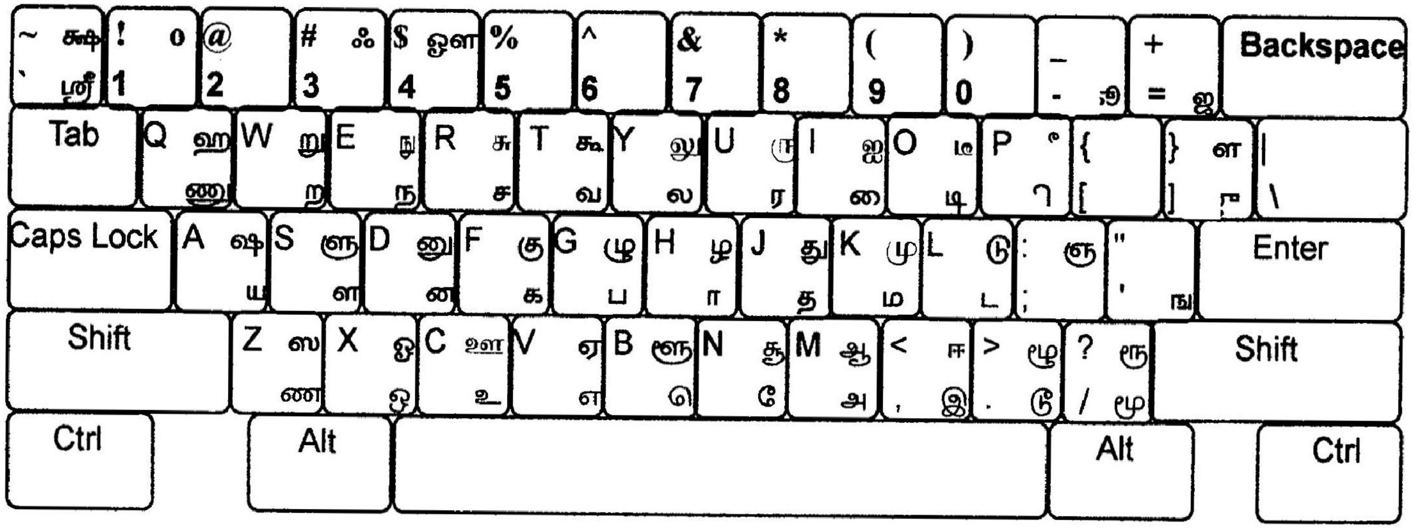



படம் - 3 தட்டச்சு வகை (மயிலை)

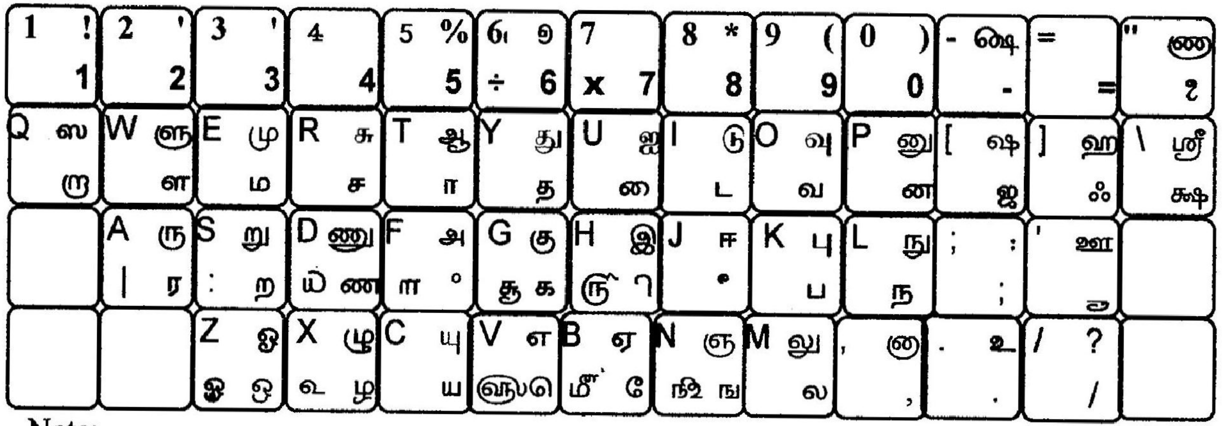
Note : ஸ்ரீ is got by typing ஷ்ர from the top row followed by °
cojuction க்ஷ can also be typed as க°ஷ

படம் - 4 **INSCRIPT OVERLAY FOR TAMIL**

படம் - 5 ***கணியன் சாவிப்பலகை***




படம் – 6 **Srilankan keyboard layout**



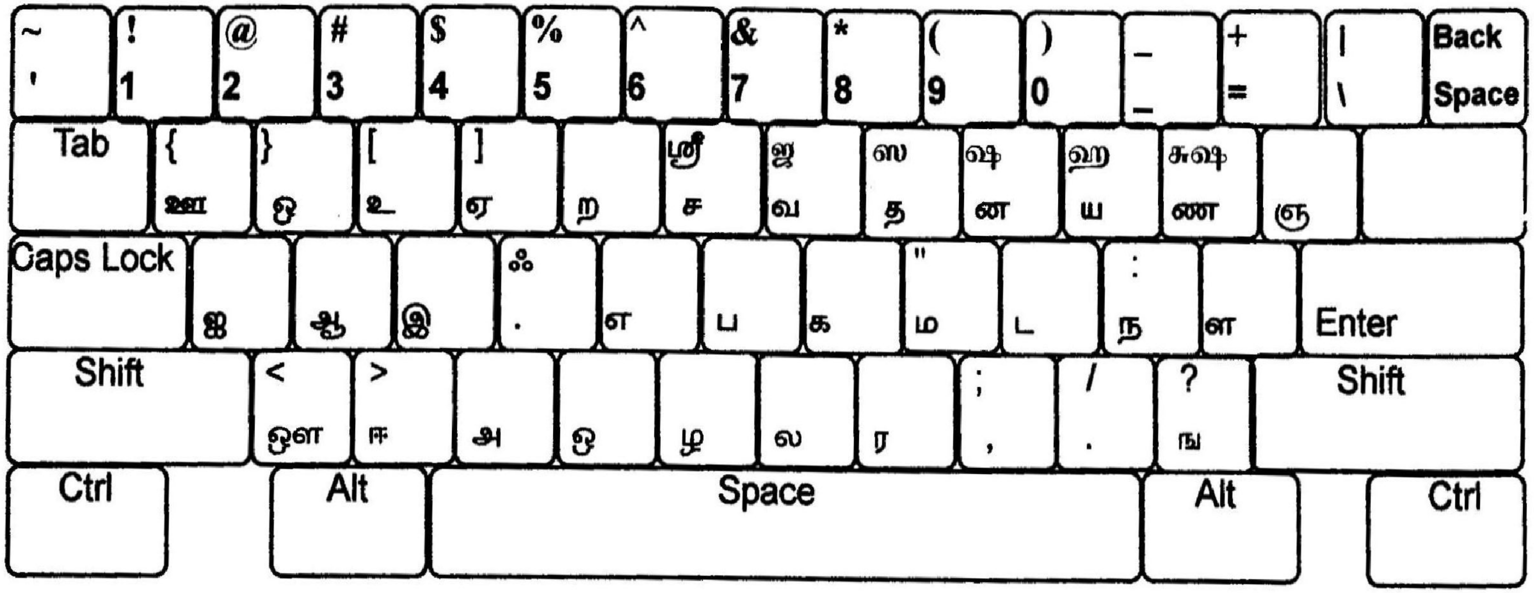
Note:

1. Alter = Control + Space (For MS Windows)
2. ஹுல மீ ஙீ Are available only in Fonts 801, 802, 803

| Key | Shift |
| --- | --- |
| Alt. | Normal |

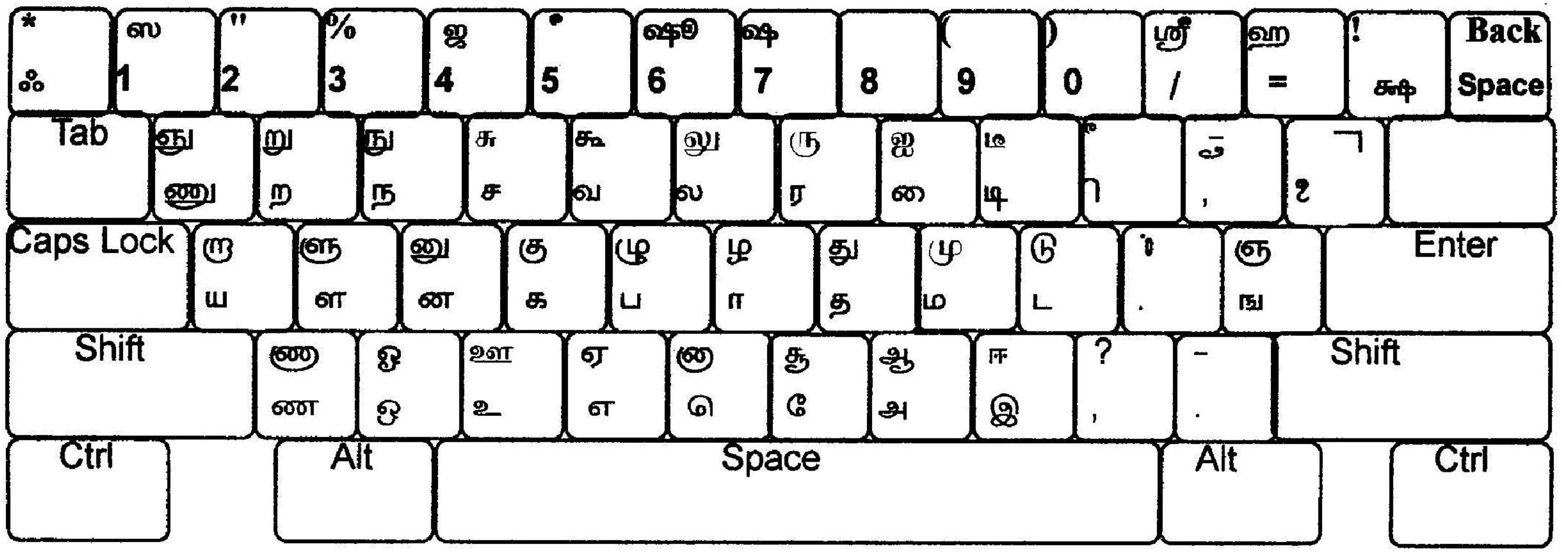
படம் - 7 தினமலர் **Krishnamoorthy layout for tamil**

படம் - 8 அறந்தை சாவிப்பலகை

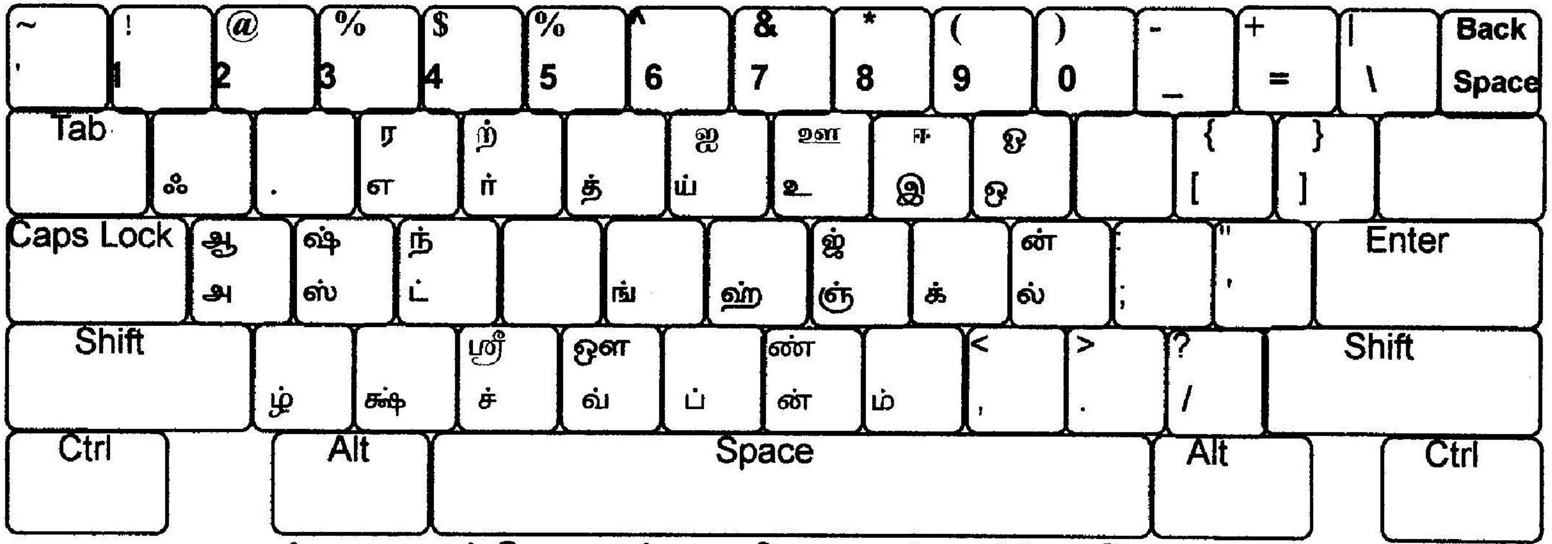



படம் - 9 ஒலியியல் விசைப்பலகை

தமிழக அரசு விசைப்பலகை தரக்குழு




படம் - 10 தட்டச்சு சார்ந்த விசைப் பலகை
தமிழக அரசு விசைப்பலகை தரக்குழு




படம் - 11 ஆங்கில எழுத்து வரிசையைச் சார்ந்த விசைப் பலகை
தமிழக அரசு விசைப்பலகை தரக்குழு

# COMPUTER KEYBOARD FOR TAMIL ISSUES INVOLVED

*Dr. V. Krishnamoorthy*
*Professor, School of Computer Science and Engineering*
*Anna University, Chennai 600 025, Tmt. Kamala Krishnamoorthy*
*Lecturer in Tamil, Govt. Arts College for Men (Nandanam), Chennai 600 035*

**Abstract**

*This article talks about the issues involved in designing an efficient computer keyboard for Tamil. A keyboard design is proposed for possible adoption as a standard keyboard. It is ergonomically designed such that the finger movements and also the number of strokes are minimized. The classification of letters as short and long vowels, and, hard and soft consonant pairs, have been effectively used to design this layout. Placing such related letters together is the basic principle used. The nature of Tamil language in favouring certain types of combinations of letters is used in minimizing the number of strokes by about 10%.*

**Introduction**

This paper is divided into four parts. They are 1. Quotations from the report of the DOE Committee for standardizing uniform keyboard for all the Indian languages. 2. The important points to note from this. 3. The need for a separate keyboard for Tamil. 4. Points to be considered in designing a Tamil keyboard.

**PART 1**

We refer to the "Report of the Committee for Standardization of Keyboard Layout for Indian Script based Computers", Electronics Information & Planning, Vol. 14, No. 1, October 1986, pp.3 - 24. This paper clearly explains the logic behind the design of the keyboard layout. The reasons behind the placement of each and every key is discussed, after considering various issues relating to the keyboard design. Some important parts from this paper are quoted below, for ready reference, as they form the basic principles in designing a keyboard in the Indian context.

On phonetic keyboards and their basic structure -

Advantages of phonetic keyboard over graphic keyboard

1. (p.6) Easy to learn since the basic vowels, consonants and matras are given in an orderly fashion, according to the inherent characteristics of Indian languages. ( The graphic keyboard is ) Difficult to learn, because of absence of all alphabets, and the random distribution of the graphic shapes on the keyboard.
2. (p.6) As an educational tool, it acquaints users with the proper phonetic nature of the Indian scripts, and the consequential similarity between them.
3. (p.6) Is very attractive for sight-typing or single-finger typing as practiced by casual users. This is possible because of the structured layout of characters according to their properties. This aspect is very important since even on English computers several users do sight-typing.

On Educational value -

5. (p.8) To be of educational value the Indian script overlay ought to reinforce the scientific nature of Indian scripts on the user.
6. (p.8) Most of the students would be sight-typing rather than touch-typing. The sight-typing becomes simple due to the structured layout.
7. (p.9) Presently, majority of word-processor users in office may not be the professional typists and as such ease of learning, simplicity of layout, soundness of approach and the efficiency of operation would be the decisive factors in their usage of Indian scripts.

Character frequency considerations -

1. (p.11) In touch-typing the ease and speed of typing increases significantly when the consecutive characters are typed by different hands. In this case the parallelism which can occur between two hands is exploited to the maximum extent.
2. (p.11) However lately it has been shown that the frequency consideration is not all that important. There are other factors which are more significant.

3. (p.12) A touch-typist is taught to stroke the keys in a uniform rhythm. Irrespective of the difficulty, the keys are to be stroked in the same time. Hence with time all key-strokes become equally easy.
4. (p.12) When the efficiency difference between the Sholes keyboard and Dvorak keyboard is so small, it leads to the conclusion that frequency considerations cannot constitute the paramount factor in evolving an optimal keyboard layout.
5. (p.12) So apart from the very broad observations, the smaller details may be misleading.
6. (p.12) In fact the English keyboard layout remains convenient for all languages, based on the Roman scripts. All this could happen since it attempted to just minimize irritation - not to optimize through frequency consideration.

Nature of letters and their placement -

1. (p.15) Actually we have three keys vacant in the bottom row of the left pad. These can take up the consonants which have to spill over to the other side.
2. (p.16) This way each varg is contained in a column, according to the frequency trend derived from phonetic considerations.
3. (p.17) The five basic vowels are given the home row. In the unshifted positions the matras are given, while the corresponding vowels are given in the shift positions.
4. (p.17) In a similar way the long forms of the 5 basic vowels are kept just above the Home row, in the corresponding positions.

## PART 2

### Important observations from the above

1. The layout should be easy to learn. This is a very important requirement. This can be achieved by using the structure of the language.
2. The frequency considerations can be a guideline and not the only deciding factor. Structured layout is more important.
3. Similarly the left-right alternation of hands can only be guideline and not an important deciding factor. For this, just see how poorly the English keyboards behave in this respect, and still remain easy to use.
4. Vowels and consonants should be given for different hands, as far as possible, to speed up touch-typing. A few consonants can be on the other side.
5. The following three types of structure of the language can be used in making the layout easy to learn. They are 1. Vowels and the corresponding matras in the same key.
   2. Short vowels in the home row and the corresponding long vowels just above them.
6. We would like to note the importance of structured layout by noting the following in the case of the Tamil typewriter keyboard.
   1. The 5 long vowels are in the upper case positions of the corresponding short vowels.
   2. As far as possible, the ukaram of a consonant occupies the upper case of that consonant.

## PART 3

### Need for a new keyboard for Tamil

With the advent of Internet and information explosion, the benefits can reach the common man only if the keyboard is very easy to use. This property of the keyboard becomes the most important one in the current scenario.

The Inscript keyboard, though a very good one for languages like Hindi, does not seem to be efficient for Tamil, for the following reasons.

1. In pure Tamil, we do not have four letters in the same varg. We have only one. We use some grandha letters sa, sha, ja, ha, ksha and sri. Due to this some lower case positions are wasted. Also this necessitates the use of some upper case positions.
2. Due to the above clustering of the four varg consonants, the natural clustering of the nasal consonants along with their natural counterparts called the strong consonants could not be accommodated. In Tamil the six strong consonants are paired with the six soft(nasal) consonants. They are called 'ina' letters. They are paired like that because they are produced from the same position of the mouth. This pairing is codified thousands of years ago in the Tamil grammar works Tholkaappiyam and Nannool. It can be seen that these pairs come together a large number of times. Also whenever they come together, normally the soft consonant comes first, and as a half consonant. The exceptions are rare. Note that the nasal consonants are created by

just one symbol Anuswar in Hindi. The positioning of this has been taken care of in the Inscript keyboard. But this is not relevant for Tamil.

3. The natural structure of Tamil consonants has not been taken into account in the Inscript keyboard. If the six pairs are kept next to each other, four benefits will occur. One is for the casual user. He/she can very easily remember twelve positions very easily. The other is for the touch typist. If adjacent fingers type these two letters, there will be no finger movement. Since such pairs come with a very high frequency, the strain on the hand will be considerably reduced. The third benefit is that, many times, the spelling mistakes can be avoided, since the occurrence of one letter implies the occurrence of its pair as the next letter. The fourth benefit is that the structure of the language is taught to the students.
4. The old mechanical typewriter keyboard for Tamil has not been popular with the casual user, primarily because of the heavy usage of the upper case positions. In the Inscript layout all the vowels and some consonants occupy the upper case positions. This is going to make things much harder for the casual user.

The above discussion shows that a keyboard layout which remedies the above deficiencies has to be designed exclusively for the use of Tamil.

The Kanian keyboard was studied next. The following points are worth noting.

1. It is designed based on three aspects. One is the frequency analysis. The next is that vowel and the consonants should be assigned for different hands. The third aspect uses one feature of the Tamil language. Each aspect is considered below, one by one.
2. The more frequent letters are assigned in the middle of the keyboard. The dot symbolizing the pure consonant is put on the key 'G'. This is the mostly used key, with a frequency of about 18%. 'G' is not the natural position in which the left index finger is kept. Hence the left index finger has to move to the right for about 18% of the time. This is an easily avoidable irritation and waste of energy. This goes against the design goal for an easy keyboard.
3. The left index finger seems to do most of the work of the left hand. The left little finger seems to do very little work. The left index finger may be doing about ten times the work of the left little finger. It is desirable that the strong fingers take a large share of the work. But the given proportion of ten times seems to be not so desirable.
4. The hard and soft consonants pair are not kept adjacent. In the following pairs, (cha, gna), (ta, nna), (tha, na), (rra, na) the finger position has to move.
5. The vowels are given to the left hand, and all the consonants to the right hand. Two consonants are put on the upper key positions. The advantage of left-right alternation may be offset by the upper case problem. Note that in Inscript keyboard some consonants are placed on the left side also.
6. In Tamil, normally a vowel does not come as a pure vowel in the middle of a word. A pure vowel comes only in the beginning of a word. There are exceptions, but they are very few. This property has been effectively used in avoiding the shift keys. Both the vowels and the corresponding matras are produced by the same lower case key. In the beginning of a word, it becomes a vowel. In the middle of a word, it becomes a matra and joins with the previous consonant. If there is no consonant to join, it becomes a vowel. Since we are talking about a computer it can keep track of the required information, and act in a context sensitive way. This improvement makes the typing easy, since the upper case positions are avoided considerably. Also it reduces the number of strokes needed since pressing the shift key is also considered as a stroke.
7. The short and long vowels are not paired, and they are kept in different places.
8. In short, the frequency analysis and the left-right alternation seems to have been literally followed. This has lead to an unstructured layout, and the loss of all the benefits arising out of it. The positive aspect is the avoidance of upper case strokes by combining the vowel and the matra in the lower case keys.

Regarding the keyboard design proposed by the keyboard committee, and published in the DinaMani dated 25 - 11- 98.

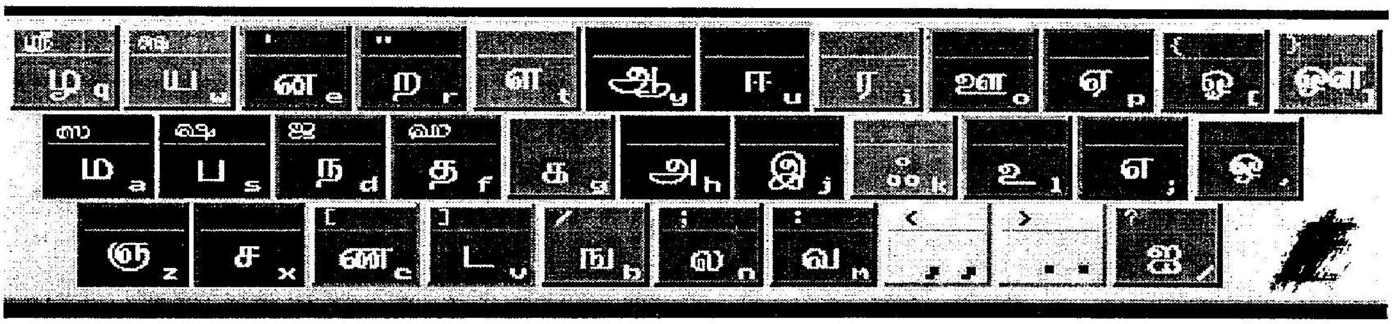
1. This keyboard is similar to the Kanian keyboard. The disadvantage of putting the dot symbol in the 'G' key has been rectified, and is put on the 'F' key. Uppercase strokes have been avoided by placing two consonants on the left side. On the left hand side the load ratio for the little and the index finger is reduced to one is to eight approximately. All the other discussions are similar.

**PART 4**

**Design of a new keyboard**

The decisions taken and the reasons behind them are given below.

1. To make it as the easiest keyboard, for all the pure Tamil letters upper case will be avoided. Only the grandha letters and some symbols will be put in the upper case.
2. In Tamil there are only 18 consonants, 12 vowels and one letter called aayutham. A vowelised consonant, which is a combined letter, called 'uirmei' in Tamil, is a combination of a consonant and a vowel. The consonant sound comes first and then the vowel sound. So it is natural that the consonant is typed first and then the vowel. It is natural to move from left to right while writing. So it will be natural to type with the left hand first and then with the right hand while forming a combined letter. Hence the consonants are assigned the left side, and the vowels the right side. Since there are 15 keys on the left and only 16 on the right this does not really make much difference. Only 3, instead of 2, consonants are put on the right side. We believe that the natural left and then right rhythm will avoid inner mental conflict. We are able to achieve this since we have only 18 consonants. Note that in the Tamil typewriter keyboard also more consonants are on the left, and more vowel/modifiers are on the right.
3. On the vowel side, the short vowels should come in the home row. The long vowels should be above the corresponding short vowels.
4. The soft consonant and the corresponding hard consonant should be adjacent, from left to right. Such pairings should be given for the (middle and index) and the (little and ring) fingers. This makes the typing easy for the fingers.
5. The frequency analysis should be used to maximize the percentage of strokes in the home row, and particularly in the home positions a, s, d, f and j, k, l, ; . Next the top row should have more strokes and finally the strokes at the bottom row should be minimized. In the bottom row also the index finger can stroke with ease. This also has to be kept in mind.
6. Note that as indicated in the Report, the frequency analysis has to be taken as a guidance and not as the most important one.
7. The enclosed picture shows the layout which accommodates the requirements as much as possible. In the case of the letter ka, its soft consonant is kept just below it, and not on its left.
8. The key 'k' acts as the key for putting a dot for making a consonant a pure consonant. When it cannot modify a consonant, the letter aayutham is formed.
9. The method of the lower case key acting as a vowel and also as a vowel modifier is used in this keyboard also.
10. The ordering of the vowels are as in the natural ordering. This will fecilitate easy remembering. It may be noted that this placement has been arrived at after careful consideration using the frequency analysis.



The definition of the keyboard can stop here, or it can be expanded still further in the following ways.

1. There are two more special properties of Tamil which can also be used to reduce the number of strokes to a considerable extent, just as above in the case of vowel/modifier.
2. We have already seen that normally when a soft consonant and then the corresponding hard consonant come, the soft consonant comes as a pure consonant. In this case it can be defined to produce the pure consonant automatically. That is by just typing a pair of soft, hard consonants, the soft consonant will automatically get the dot. Note that recognizing the soft and hard consonant pairs is going to be easy, since these pairs are conveniently put next to each other. This action reinforces the pairing concept in the student.

3. Similarly, in Tamil, when the same consonant comes twice adjacently, mostly the first will be a pure consonant with the dot sign. Here again it can be defined that, typing the same consonant twice in succession, will automatically make the first occurrence as a pure consonant with the dot sign.
4. There are exceptions for the above two rules. If the first consonant should not become a pure consonant, then it will be a consonant with the vowel 'a'. To specify this, it is defined that when 'a' is typed after a consonant, then that consonant is confirmed as a full consonant, and it will not join with the next stroke.
5. What happens when a soft consonant is followed by its corresponding hard consonant twice? The soft consonant gets a dot first. Another rule can put a dot on the middle hard consonant. But in Tamil, the grammarians have coded when two half consonants can come one after another. They are only ya, ra, zha followed by ka, cha, tha, pa, gna, na, ma. Since ya, ra, zha are not soft consonants, the following rule can be adopted - When a sot consonant is followed by its corresponding hard consonant twice, only the soft consonant becomes a half consonant. The first hard consonant remains as a full consonant. Similarly, if the same consonant is pressed three times consecutively then the first one will become a half consonant, and the second will remain as a full consonant.
6. These above definitions are easy to implement in a computer, just as in the case of context sensitive vowel. This improvement reduces the number of strokes needed to type a passage by about 10%. Think of the energy savings in the global context ! Note that we are talking about designing a keyboard for a COMPUTER. We are supposed to use the structure of the language and also the capabilities of the computer to make the keyboard operation easy and efficient.
7. The discussion on the design of a computer keyboard is not yet over. There are other issues to be decided befor implementing any keyboard design. If there is uniformity in these features also, it will greatly help the users. First, what happens when the current cursor position is changed by the movement of the mouse? What happens when the backspace, delete key and the arrow keys are pressed? Note that these questions have very simple answers in the case of English. But in the case of Tamil, since the combining of letters are involved, they may have to be decided properly, for the sake of uniformity. Also it may be noted that each type of keyboard may require a different answer for these qestions. In this context, the implementations in the other systems like DOE keyboard have to be studied befor taking a decision.
8. In the case of a changed position of the cursor, the computer can sense this and then (1) Not join with the letter just before the cursor, or, (2) Join with the letter just before the cursor if possible.
9. In the case of backspace there are three choices. In Tamil letters like kow and ko are made up of three symbols. Normally the backspace erases one symbol. In the case of the phonemic keyboard such letters are created by two strokes, one for the consonant and the other for the vowel. It is natural to expect that the backspace will erase the last stroke. This means that ko will be erased and ka will appear. This will be very helpful when a vowel key is pressed wrongly, and it is to be corrected. Erasing a stroke makes the correction possible in minimum number of strokes. For example ko can be corrected to ki with just two strokes, backspace and i. This method is natural and efficient. There is a third method. In this the entire letter can be erased. But this is not natural, and also takes more strokes for corrections. For example in this method, three strokes are needed to correct a ko into ki.
10. In the case of delete, deleting a stroke does not make sense since from a vowelised consonant a vowel will be remaining. An existing vowel does not join with a consonant. Hence this will create hanging vowels. More strokes will be needed to delete them. Hence there are two options for the delete key.
11. In the case of arrow keys the cursor may move one symbol at a time or one letter at a time. In the case of phonetic keyboard, it may be better if the arrow keys move on letter at a time, and in the case of the typewriter keyboard it may be better if it moves one symbol at a time.
12. To distinguish the keyboard defined here from the others, it has been named as Arandhai keyboard. The approximate percentage of load to the different keys and different fingers are given below. This shows a desirable distribution of load.

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Q:1.1 | W:2.8 | E:4.4 | R:3 | T:2 | Y:4 | U:0.6 | I:4.2 | O:0.6 | P:1.4 | { 1 | } 0 |
| A:5.4 | S: 4 | D:2.4 | F:6.7 | G:7 | H:1.5 | J:7.8 | K:9.5 | L:9.3 | ; 2 | '1.3 | |
| Z: 0.1 | X:1.8 | C:1.5 | V: 3.5 | B: 0.6 | N: 3.7 | M:3.7 | | | / 3.1 | | |

| | Left hand | Right hand | Total |
|---|---|---|---|
| Index finger | 22.8 | 21.3 | |
| Middle finger | 8.3 | 13.7 | |
| Ring finger | 8.6 | 9.9 | |
| Little finger | 6.6 | 8.8 | |
| | | | |
| Top row | 13.3 | 11.8 | 25.1 |
| Home row | 25.5 | 31.4 | 56.9 |
| Bottom row | 7.5 | 10.5 | 18 |
| **Total** | 46.3 | 53.7 | 100 |

It shows that the left hand is given a lighter load than the right. Also the home row itself takes the bulk of the load, followed by the top row.

1. An earlier version of this keyboard has been implemented in a word-processor named Padhami, a few months ago. This word-processor is being received well by the public. The Arandhai keyboard layout has been successfully implemented in a word-processor named PonMozhi also recently.
2. In PonMozhi regarding the cursor, backspace and delete keys, all the possible options are provided as an experimental measure. The user can choose any one. It seems that sensing the context and joining if possible may be a better choice, for a changed cursor position. In the case of the backspace removing a stroke seems to make typing easy. In the case of delete not much difference is noted between the two options. This is from the personal experience of the author. In the case of the arrow keys, moving one symbol at a time is implemented.
3. Tamil letters have been accommodated without using the shift keys. Many of the shift positions are vacant. These shift positions can even be used to type symbols like multiplication, division, square root, degree, tick mark, bullet, one-fourth, half etc.. This will make the typing of Tamil text with mathematical symbols far easier than typing the English text with mathematical symbols. Currently, since these symbols are not available on the English and Inscript keyboards since these keyboards use all the upper case positions. They have to be inserted as symbols or by other methods.

## CONCLUSION

It is argued that by learning one Inscript keyboard one can type in any Indian language. Yes. It is true. But how many such typists we really have or need? It has been clearly made out in the Report that a vast majority is going to be casual users, and , their needs have to be taken care of. We have seen that this majority is going to increase more and more because of the usage of Internet. We have shown that the new keyboard design excels in most of the design goals set in the Report, of course for the Tamil language only. The main thrust being in the areas of easiness and the usage of structure of the language. This is because we optimized for only one language, whereas in Inscript they had to cater to many languages. We do not underestimate the relevance of the Inscript keyboard. We have started with the design goals as set in that Report. This Report has been the starting point for the design of the new keyboard. The new keyboard design saves around 10% of time and energy. All these reasons strongly recommend that this new keyboard definition, if adopted as the standard, will bring in the benefits of the computer revolution to the rural population also, and also save an enormous amount of scarce resource available in our country.

# TAMIL IN CYBERSPACE

*'jAzan' Ramalingam Shanmugalingam (Appu Archie)*
*36, Farrell Court, Marblehead, MA 01945 (USA)*

Courage is the keyword that comes to mind as one ponders about the rapid changes that are taking place before our eyes, and ears. One needs courage to face new situations that at the outset may be perplexing and signal deep changes in our ways of life. Cyberspace to put it simply is, "The virtual space created by Computer Systems." There are many numbers of questions that may arise out of making this world into a "Global Village." Until such lofty ideas take any shape, life has to go on and we have to find ways to overcome fears of new challenges and opportunities that will serve to make our lives better. Tamil, more than anything-else, is the mother tongue of more than 70 M people. The tradition and culture of the Tamils allow them to mix and assimilate easily with other ethnic groups and yet maintain their identity in tact in most cases. Absence of a free Tamil country has begun to show signs of loss of Tamil identity. Nevertheless, the enthusiasm and interest in Tamil shown by a few Tamil expatriates in the so called Developed countries is a good omen to future retention of Tamil interest via Cyberspace as well.

It becomes necessary to ignore the realities of most Tamil lives in developing countries, as economic conditions will dictate a more somber attitude towards investment in more sophisticated technological advances when there is still the hang-ups about basic human needs. Thus hampering much needed technological advances. In the meantime, we should not feel increasingly isolated, dependent on others, apparently without needed resources to find a lifestyle we deserve. The days of a rigid agrarian society are a distant memory and happily so, even though a larger population is dependent on Agri based industries. Technology proved to be a major gift in changing the rigid society in the last 50 years into a society open to its promise and given the environment, learn well enough to use technology properly. But unlike the Americans. Tamils should not replace a solid based "hardware" industrial complex with a more abstract cyberspaced virtual reality type of bubble economy. It is said that, that the most valuable corporation in the US and the world is Microsoft, the software company. If the world economy is dominated by 'information' companies, financial shocks are bound to happen and this may lead to the inevitable "Big church" some scientists predict will happen sooner than later. In short we should develop our own software that we can use to make our life better but not to spell disaster by depending on others.

Computers can break through barriers in life and connect with other people. Computers are beginning to bring the whole to us. It is true that many of us are threatened by computers and all the mysterious things connected with it. The car is a complicated piece of equipment and we are not threatened by its ability to function with the least effort and knowledge of it by us.

Yet its dangers on the road are not minimized by our affinity for the car. Similarly, the computer in addition to being a hobby can also cut down distance and time to bring people closer in spirit through esprit de corps. The computer can send messages in numbers unlike the telephone, where it is a one-on-one mode of communication. Today, we can face each other though in virtual reality as photo-images, discuss and decide matters as if we are physically present together. Also unlike TV with its limited channels, time and distance of access, the INTERNET is almost an endless ocean of data in Cyberspace, where much of the world's information can be found.

I am no "Technophile," nor a "neophyte," on computers. I am learning to be an expert in the use of a computer, and leave the web of designing and developing the more complicated technology to experts. There still are some important useful people in the USA who claim to live out their days "Computer-free." I have found the computer to be a source of inspiration to learn more, communicate better, and less strain on the purse. It will be even more economical, if its use is widespread and in native languages. It is this desire that prompted me to get involved in the design and development of a computer tool for Tamil.

I was disappointed in the many English translations of THIRUKKURAL Most translations appeared to me as either paraphrases or non-couplets "English" according to Richard Lederer, in his

book "Miracle of English as a preferred second language outnumbers those who use it natively. English is first or official language of 45 countries." Thirukkural, in its native Tamil is unmatched for "brevity, the soul of wit." Yet, English with its large vocabulary, burrowed or added to its "dark age" low Germanic tongue in the mid 5th century, through the Anglo Saxon Old English until William Duke of Normandy led Normans invade England in 1066, conquered Saxons and Danes. Duke William became King of England. Free Mix of French words into Old English gave way to Middle English. "European renaissance between the 14th to 17th centuries began as educated people rediscovered the world of ancient Greece and Rome. Love affair for anything classical led to taking of Latin and Greek words that could describe new discoveries in Medicine, Art, Science, Geography, etc., These classical words entered into English through writing-often scholarly writings."

It is estimated, that there are more than one million words in the English language although only about 615,000 have been listed in the Oxford English Dictionary. Poet Carl Sandberg said, "The English Language hasn't got where it is by being pure" Naturally. I began to wonder, why, that THIRUKKURAL that can claim to biblical usefulness, has no English version in couplet form e.g.

> *"perukkattu vEnhdum panhital; ciRija
> curukkattu vEnhdum ujarvu." KuRaL 963
>
> *பெருக்கத்து வேண்டும் பணிதல்; சிறிய*
> *சுருக்கத்து வேண்டும் உயர்வு.* *குறள் 963*
>
> Humility in Prosperity
> Dignity in Adversity. English by Appu Archie

*I am using and will use Yarzhan Key to pronunciation when using English Phonemes in all my writings. An Alphabet Chart in Roman Script for Tamil is given below:

[illegible] KEY TO Pronunciation Using Roman Script
By AppuArchie

| [illegible] | | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] | [illegible] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | a | A | i | I | u | U | e | E | y | o | O | w |
| [illegible] | k | ka | kA | ki | kI | ku | kU | ke | kE | ky | ko | kO | kw |
| [illegible] | ng | nga | ngA | ngi | ngI | ngu | ngU | nge | ngE | ngy | ngo | ngO | ngw |
| [illegible] | c | ca | cA | ci | cI | cu | cU | ce | cE | cy | co | cO | cw |
| [illegible] | gn | gna | gnA | gni | gnI | gnu | gnU | gne | gnE | gny | gno | gnO | gnw |
| [illegible] | d | da | dA | di | dI | du | dU | de | dE | dy | do | dO | dw |
| [illegible] | nh | nha | nhA | nhi | nhI | nhu | nhU | nhe | nhE | nhy | nho | nhO | nhw |
| [illegible] | t | ta | tA | ti | tI | tu | tU | te | tE | ty | to | tO | tw |
| [illegible] | N | Na | NA | Ni | NI | Nu | NU | Ne | NE | NY | No | NO | Nw |
| [illegible] | p | pa | pA | pi | pI | pu | pU | pe | pE | py | po | pO | pw |
| [illegible] | m | ma | mA | mi | mI | mu | mU | me | mE | my | mo | mO | mw |
| [illegible] | j | ja | jA | ji | jI | ju | jU | je | jE | jy | jo | jO | jw |
| [illegible] | r | ra | rA | ri | rI | ru | rU | re | rE | ry | ro | rO | rw |
| [illegible] | l | la | lA | li | lI | lu | lU | le | lE | ly | lo | lO | lw |
| [illegible] | v | va | vA | vi | vI | vu | vU | ve | vE | vy | vo | vO | vw |
| [illegible] | z | za | zA | zi | zI | zu | zU | ze | zE | zy | zo | zO | zw |
| [illegible] | L | La | LA | Li | LI | Lu | LU | Le | LE | Ly | Lo | LO | Lw |
| [illegible] | R | Ra | RA | Ri | RI | Ru | RU | Re | RE | Ry | Ro | RO | Rw |
| [illegible] | n | na | nA | ni | nI | nu | nU | ne | nE | ny | no | nO | nw |
| [illegible] | J | Ja | JA | Ji | JI | Ju | JU | Je | JE | Jy | Jo | JO | JW |
| [illegible] | S | Sa | SA | Si | SI | Su | SU | Se | SE | Sy | So | SO | Sw |

vanhakkam

Your comments with suggestions if any welcome Thanks Shan

I had difficulty typewriting Tamil in computers before 1993. Whatever software or Fonts available for Tamil then were based on the Typewriter Keyboard. One look at the Keyboard Matrix convinced me that there must be a better way to make Tamil typewriting with computers easy and logical, a prerequisite for any Tamil software. My engineering background and love for my mother tongue prompted me to look for a Tamil answer to a word processing problem in Tamil depending on a Western or English origin computer for Tamil. Every language has its share of Literature, Music and Drama, but Tamil is perhaps the only language that has given language a life and body, namely 'ujir and mej' and all Tamil letters are either life or body and a combination of body and life give body with life letters or 'ujirmej ezuttukkaL.'

Tamil has 247 letters, depending on 31 Basic Letters thus:

| | | |
|---|---|---|
| Vowels - 'ujir', 'aA, iI,, uU, eE, y, oO, w(av) | = | 12 |
| (Of the 12 'ujir' 5 are short and 7 are long vowels.) | | |
| Velar Fricative - (∴) | = | 1 |
| Consonants - 'mej' | = | 18 |
| Total Basic Letters | = | 31 |

Each Consonant in combination with each Vowel can produce a third letter or a dependent letter called 'ujirmej.'

∴ 18 Consonants and 12 Vowels produce Dependencies - $18 \times 12 = 216$

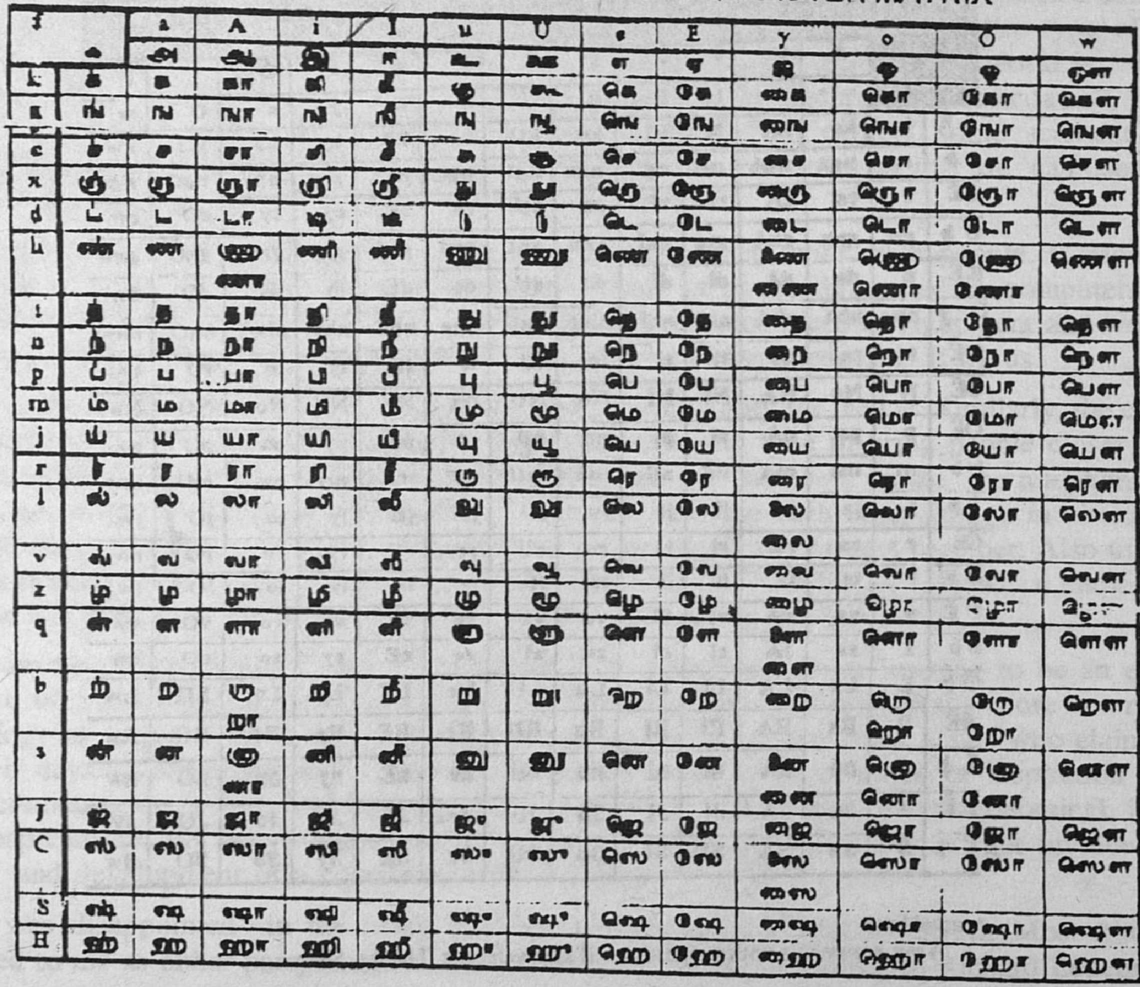
∴ Total number of Tamil Letters =247

The 5 pairs of short and long vowels can be represented by 5 upper and lower cases of the QWERTY Keyboard.

∴ Number of Keys required to represent 31 basic letters is (31-5) = s26

Pleases refer to the CPD- YARZHAN TAMIL-EDITOR CHARACTER MATRIX below:

CPD YARZHAN TAMIL EDITOR - CHARACTER MATRIX



Obviously, Tamil language with its renown antiquity falls easily into modernity by being easily represented by the 26 Keys of the QWERTY Keyboard without exception. There is one shortcoming in using all 26 Keys to represent Tamil, as any transliteration will not be accurate, since letters represented by sounds not in Tamil will give rise to confusion.

*SECTION - II*

# Standardisation of Coding of Tamil Scripts

# TOWARDS A TOTAL INTERNET SOLUTION FOR THE TAMIL LANGUAGE THROUGH SINGAPORE RESEARCH

*Naa Govindasamy*
*Lecturer, National Institute of Education (NIE)*
*Nanyang Technological University (NTU)*
*469, Bukit Timah Road, Singapore 259756*
*http://www.irdu.nus.edu.sg/tamilweb*
*email: govind@cir.nus.edu.sg*

(This Paper was presented at the SAARC Conference on Extending the Use of Multilingual & Multimedia Information Technology at Pune, India, on September 1 –4 1998. The present paper has added new developments which took place after September 1998. )

(The author wish to thank Dr Tan Tin Wee, Associate Director for the Centre for Internet Research (CIR), National University of Singapore and Mr Leong Kok Yong, Research Officer of CIR, for giving valuable advice in the preparation of this paper.)

**Abstract**

*(Internet revolution has enabled the widespread dissemination of information throughout the world. Most of the content is in Romanized characters. Research is going on in some countries to enable non-Roman scripts accessible on the Internet. This paper will discuss and demonstrate how, through a successful research collaboration in Singapore, Tamil language content is now freely accessible, searchable, conveniently emailable and easily composed and edited on the Internet through all three popular platforms Unix, PC and Mac. The important part of this paper is latest development, which took place in Singapore after September 1998.*

1. Multilingual Text Editor For Unix, Windows and Mac. Text is saved in Unicode and Tamilnet code. This is the first Unicode Tamil Editor to be seen on the Internet for free downloading.
2. Multilingual multiscript URL is another important development in Singapore Research. My presentation will focus on these two important delopment.)

## 1. The Need for A Multilingual Internet in Singapore

Singapore is a multilingual and a multiracial country. English, Chinese, Malay and Tamil are the official languages. Most of the government and public documents are published in these four official languages. However, until recently, it was not possible for the Chinese and Tamil languages to be disseminated through the World Wide Web on the Internet. In 1994 Dr Tan Tin Wee, my research collaborator initiated work in this area while he was the head of Technet Unit, the first Internet service provider for Research and Educational Institutions in Singapore. Technet Unit was directly under the supervision of the Computer Centre of the National University of Singapore. (Technet Unit has since been commercialised to become Pacific Internet, one of the three ISPs for Singapore. The others are Singnet and Cyberway.)

By mid-1994, Technet Unit initiated the Singapore INFOMAP project which provided a one-stop WWW home page for Singapore. He wanted the four official languages to be represented in the INFOMAP. Since English and Malay are using the Roman script, displaying these two languages on the WEB was not a problem. By the end of 1994, Technet had successfully implemented an Experimental Chinese WEB server in Singapore. So the problem of displaying the Chinese script on the Web was solved.

However, displaying Tamil script on the Web, and communicating through Tamil on Internet was a problem.

There was no Tamil Information System on the Internet which provides a display system in Tamil and English simultaneously on the Text Mode using a Tamil-English single font file. There were a few servers, which were providing Tamil script using GIF image files. Tamil Eelam Page (http://www.eelam.com) was and is still very active in this direction. Tamil Nadu Home Page, and Tamil Electronic Library .

(http://www.geocities.com/Athens/5180/index.html) are other popular Tamil Web Sites on Internet at that time. Tamil Electronic Library was using (and is still using) a mono 7bit font (Mylai) for the Tamil display on the Web.However, Mylai font cannot support native emailing at that time. So there was a need to develop a Tamil Internet System which should go beyond Web display.

In May 1995, I met Dr Tan at the Technet Unit, National University of Singapore, soon to become the Internet Research and Development Unit (IRDU) (now upgraded to Centre for Internet Research - http://www.cir.nus.edu.sg). We identified the potential solutions and agreed for a possible research collaboration between NUS and my institution, NIE, NTU, the two instititutions of higher learning in Singapore at that time. (http://www.irdu.nus.edu.sg ) & (http://www.nie.ac.sg )

At that time, on my own, I was in the process of developing True Type fonts and a Tamil software for Windows Applications (Kanian™ Tamil Software).

As the service provider arm of the Technet was sold to a private consortium and renamed Pacific Internet in September 1995, the nascentTamil Internet Research was inherited by the newly formed Internet Research and Development Unit (IRDU). This Research Unit was funded by the National Science and Technology Board (NSTB ) (http://www.nstb.gov.sg ) of Singapore. Mr Leong Kok Yong, just graduated from the Nanyang Technological University, joined IRDU, and became one of the key member for the TamilWEB project.

## 2. Objective

During the Technet period, when Dr Tan and I, conceptulised the TamilWEB (http://www.irdu.nus.edu.sg/tamilweb) Project. We had a very clear objective. That is:

1. ***to develop a bilingual font system for The Total Internet Communication in the Tamil language.***

That means, the system that we intended to develop:

*a)* ***should provide display of Tamil & English Text simultaneously on the Internet Applications. (Web Browsers & Email sofware packages.)***

*b)* ***Tamil and English text should be easily communicated and retrieved in Plain Text.***

*c)* ***should work across Platforms. (PC, Mac & Unix)***

*d)* ***should be searchable in Tamil***

*e)* ***should let the user type Tamil in the Web browser's Forms, and the typed word should be seen in Tamil.***

*f)* ***Should allow the user to read Tamil & English in terminal emulation mode (telnet).***

## 3. Prototype Testing, Preview and Official launch

The prototype of our system was tested during the launch of PoemWEB. PoemWeb (http://www.irdu.nus.edu.sg/poem ) is an electronic selection of representative poems from the four official languages, from the book , Journeys: Words, Home and Nation - Anthology of Singapore Poetry (1984-1995) which was published by The Centre for the Arts, National University of Singapore. This book was launched by H.E. Mr Ong Teng Cheong, President, Republic of Singapore on Friday 27 Oct 1995.

The preview of the first phase of TamilWEB project was shown to the public and to the Press at the "Internet for Everyone – 1995" at Suntec City Exibition Hall during 1 –3 December 1995.

TamilWeb was officially launched by the Honourable Member of Parliament, Dr Ong Chit Chung (Chairman, GPC for Education and MP for Bukit Batok) on 2 February 1996. (http://www.irdu.nus.edu.sg/tamilweb)

Since then the Tamil language teachers in Singapore and the Internet Users from locally and abroad, are using the Singapore System to communicate in Tamil & English over Internet and have created a significant volume of bilingual Web pages in Tamil and English.

One of the most important Web site for teaching and learning of Tamil language was created by University of Pennsylvania's Penn Language Center (PLC).

(http://www.ccat.sas.upenn.edu/plc/tamilweb/tamil.html) Using our system, the Center created a bilingual Website for Learning and Teaching Tamil in 1996. The project is funded partially by a grant from the Consortium for Language Teaching and Learning, with the joint participation of Tamil-teaching faculty at the Universities of Chicago, Cornell, and Pennsylvania.

The ruling party of Tamil Nadu Dravida Munneetra Kazhagam (DMK) Website is another important site, using our system. (http://www.thedmk.org)

## 4. Purpose of this paper

This paper will try to explain and demonstrate, how Tamil, one of the Indian languages, has achieved the Total Internet Solution, through Singapore Research. Most of my presentation will be done through Internet.

When Dr Tan and I conceptualised the TamilWEB project in 1994, there were only two major graphical Web browsers on the Internet, Mosaic and Netscape2. Eudora was the one of the most preferred graphical Email program. PINE was the most preferred, Text-mode Unix Emailer, and LYNX was the Textual Browser on the Unix platform. Font Tagging was not available at that time. Content on the Web was almost exclusively in Romanized characters. English language was dominating the Web. Content in non-Roman languages was limited. With these background, we delivered these Internet Tools for our Tamil users.

## 5. Tamil Internet tools from Singapore

These Internet Tools originating from our research and software development are free for downloading:

1. *TamilNet.ttf (PC propotional font)*
2. *TamilFix.ttf (PC fix width font)*
3. *Tamilnet.hqx (Mac propotional font)*
4. *Tamilfix.hqx (Mac fix width font)*
5. *Tamilnet18.bdf (Unix font)*
6. *Tamilfix.bdf (Unix Font)*
7. *Tamil Keyboard Manager (for PC)*
8. *Tamil Keyboard Manager (for Mac)*
9. *Xkeymap (Tamil input system for UNIX)*
10. *Mirage (CGI Application Software for rendering Multilingual encoding text into GIF images for display on web browser*
11. *Applet input sysytem for Tamil word search*

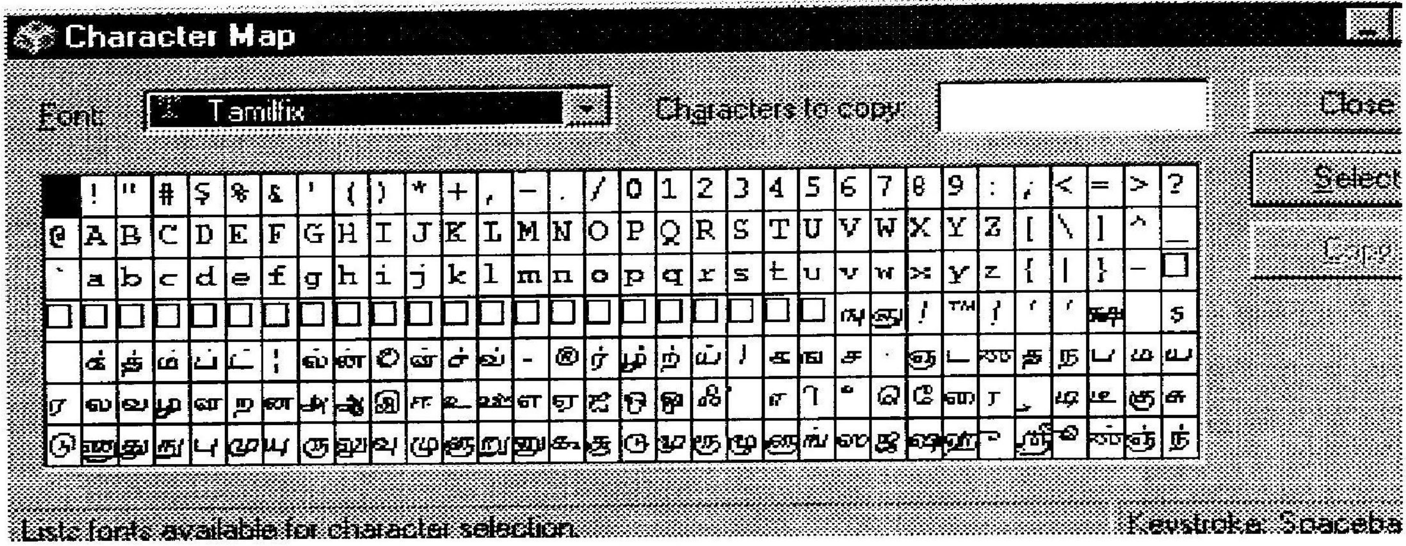
For more information, please consult the URL - http://www.cir.nus.edu.sg/tamilweb

## 6. Font encoding

The key tool for the project was the creation and design of a Bilingual Tamil-English single font system.

We have designed a bilingual font set for the display of both Tamil and English simultaneously. This was done by making use of the upper extended ASCII character range for the Tamil characters, while retaining the basic English alphabet and punctuation intact in the lower ASCII range. This will allow most of the Web world in English (or other Romanized languages) to be traversed; at the same time, Tamil codes will be recognized and displayed correctly when they occur, without having to change font set. Figure below shows the character map for the Tamil-English font set.

One important point to note is that the upper ASCII portion does not have enough code space




to include all the possible Tamil character glyphs (>200). As such, we made use of the kerning feature built into the Postscript and the True-Type font technology to combine two Tamil characters into a new character glyph not found in the above ASCII table. With the combination of two simpler character glyphs to give a more complex glyph, we can then include the entire Tamil character set within one single font, together with the English alphabet. To allow users to input these Tamil characters, a corresponding keyboard layout mapping has been devised by mapping the keys on a normal English (QWERTY) keyboard to the extended ASCII range where these Tamil characters reside. A toggle key enables the user to switch between the two modes.

"Tamilnet" propotional font was developed to display Tamil & English on the Web browser. However, the variable proportional font cannot be viewed in the Web browser's Forms.

For this purpose, a fixed width font, "Tamilfix" was developed. This font is very similar to the "Courier" font. The "Tamilfix" font is simply doing the work of the "Courier" font. Only with this fixed width font, Tamil can be typed in the Web browser Forms. In the Web browser, the form filling feature is a very important component for interactivity. If the user wish to communicate in Tamil to webmaster or the author of the webpage, he or she has to type Tamil into the Form.

## 7. Keying in the Tamil characters in the Forms

When the user is keying in Tamil script in the Form, the Tamil Characters should appear on the Forms as they are typed for immediate visual feedback. Only then can meaningful communication and interactivity take place. We achieved this through the "Tamilfix" font and the keyboard input system.

Another important factor in any database creation is the Search function. If a Search function is not possible in a particular system, creating a database, is out of the question. When we launched the TamilWEB on 2.2.1996, we demonstrated the Search function. In the search form, Tamil words were keyed in for searching against a database of Tamil text.

For search and retrieval, the submitted string in extended ASCII for Tamil (and in English as well for bilingual searches) is parsed by the httpd server and submitted as a search string to any indexing engine that has multilingual capability. In the case of Tamil, we used a simple WAIS-SF indexer and demonstrated the utility. Hits were returned in the same encoding, and displayed in the same way as described above, with bilingual capability. In fact, this powerful search function is taking place across the various platforms.

In the Singapore Government Web site (htt://www.gov.sg), searching for Tamil keywords by typing Tamil script is possible. The I Agent (http://iagent.iti.gov.sg) search engine will deliver the results in the form of webpages, using Singapore Tamil font encoding.

However, in some cases, users are unable to use our fonts and encoding system for unknown reasons. In this situation, we have invented another solution.

Our research team has produced a CGI Application software for rendering multilingual encoding text directly into images for display on any web browser as embedded images. It is called Mirage. (http://www.cir.nus.edu.sg/multilingual/mirage) When this application is added on to the server, the server is capable of rendering Unicode Tamiltext into images for the client browsers, without any helper application or any font installation. The significance of this system is that, in the client browser, the user should be able to view multilingual information, originally coded using Unicode.

For that matter, any encoding can also be transformed into images using the MIRAGE system, eg. Unicode, ISCII, Kanian-Tamilnet etc. simply by modifying the code table mappings to character glyphs.

Now, I will be demonstrating another of the important feature of our system.

## 8. Viewing Tamil on PC Terminal Emulation

When we developed the "Tamilfix" font, we knew that it will make Tamil readable in the PC Terminal emulation (eg Telnet). A Shell access user can read Tamil text in the WWW textual browser "LYNX". He can also read Tamil in the Terminal Email software "PINE".

This is a very important development for the Tamil language. In many developing countries, the number of SHELL access users typically outnumbers the TCP account users. Most users access the internet through a character-based terminal emulation rather than a graphical user interface. As such, our system benefits a lot of SHELL access Internet users. This was made possible with our "Tamilfix" font.

## 9. Keyboard Input Systems

Based on the Phonetic system, a phonetic keyboard for the Hindi language was developed by Mohan Thambi in India in 1983. This was subsequently adopted as the main keyboard for the Indian languages by the Department of Electronics (DOE) of India. However, this keyboard is based on Devanagiri script. Since Tamil is from the Dravidian language group rather than the

Indo-Aryan, of which numerous other Indian languages belong (e.g., Hindi, Marathi, Punjabi, Bengali) (Grimes, 1992), the DOE keyboard is not particularly suitable for the Tamil language.

To overcome the keyboard problem for the Tamil language, the author began an investigation into the frequency of occurrence of the Tamil vowels and consonants used within the language. Based on this research, a Tamil phonetic keyboard layout was introduced for Tamil computing (Govindasamy, 1989), named the Singapore Tamil Keyboard (Govindasamy, 1994a). Later in September 1994, the name was changed to Kanian Keyboard at the First Computer-Tamil Conference at the Anna University, Madras, India (Govindasamy, 1994b).

The keyboard consists of the 12 basic Tamil vowels placed on the left-hand side of the keyboard and the 18 consonants. The 28 basic vowels and consonants are placed in the lower case of the keyboard, while the 2 least frequently occurring Tamil consonants are placed at the upper case with the 5 Sanskrit sound consonants. For modern Tamil, a vowel will not appear in the middle or at the end of a Tamil word; it will appear only at the beginning of a word. These basic rules were taken into account when this keyboard layout was designed. The advantage of this keyboard layout is that 99.5 percent of the time, Tamil characters can be typed without pressing the shift key at all. Moreover, the most frequently used vowels and the consonants are placed at the home keys (the middle row of the keyboard). This allows the user to type 68 percent of the Tamil words by using only home keys.

Because of its simplicity and the incorporation of the Tamil grammar, this keyboard layout is very popular in Singapore and Malaysia and has been incorporated in numerous Tamil front-end processing software and word processors, including a commercial version available from the author (Govindasamy, Kanian Bilingual Wordproceesor for PC, Mac & Internet).

<Kanian@kanian.com>

The project team has devised a Tamil type writer version for PC & Mac, using KeyMan Keyboard Manager. From our server we are using KeyMan keyboard Manager. For Unix Xkeymap was used to develop the Keyboard Manager for Tamil. These three input devises are downloadable from our Website.

### *9.1 Toward Java keyboard input systems for total cross platform compatibility using Unicode.*

Java Input Method Engine (JIME) for Java from CIR, bundled native input methods and character display support in a set of applets. (http://www.cir.nus.edu.sg/jime) At present the users can input Chinese, Japanese or Korean text in HTML form irregardless of the locale of the user platform.

The Tamilweb project team has developed the Multilingual JIMEWORD. Text editor.

*JIMEWord* - JIMEWord is a multilingual WYSIWYG text editor. It supports Chinese, Japanese, Korean, French, German, Thai, Cyrillic, Greek and Tamil. It's implemented in Java 1.1 and run on all JDK 1.1 compliant platform. It runs on Unix, Windows95/NT and Mac. Able to edit, load and save text in Unicode UTF-8 and UTF-7 encoding as well as other native encoding like GB2312, BIG5, JIS, KSC, TIS and Tamilnet-Kanian (Singapore Tamil encoding). (Now this Java Input Engine is used in Yahoo Chinese Search - http://www.gime.com/ ) As for Tamil, this is the first Text editor, which runs on all platforms and saves the text in the Unicode encoding. (NT5 which is supporting Tamil and Devanagiri is yet to come on the market.) Users can download the software at: http://www.cir.nus.edu.sg/jime/

## 10. Future Direction and Conclusion

### *10.1 Multilingual multiscript URL*

Today, the Internet has reached the four corners of the world to a diverse community with different languages and cultures. The World Wide Web has progressed to address the localization needs of its audience with Web pages in different languages a reality today. However, the Internet Domain Name System (DNS) which started out to be strictly based on a subset of the Latin 1 alphabet, is still mainly English. This restriction also applies to other aspects of the Internet which makes use of domain names as well, e.g. telnet, ftp,email, etc. Now CIR is creating an experimental internationalized DNS as proof of concept that it is viable. It is also creating tools and applications that will enable users to key in URLs in multilingual characters (e.g. Chinese, Japanese, Korean, Tamil etc) It is also designing a URL forwarding system for multilingual-character URLs.

When we achieve multilingual directory and filenames, we will have fully delivered a Tamil script URL in addition to Tamil content on the internet. For the proof of concept visit http://www.idns.apng.org

Now TamilWeb Project is slowly moving to Unicode text archive. The whole Thirukural and part of Purananuru are in two coding in our server. One in Singapore "kanian/tamilnet" coding. The next coding is the Unicode.

In future most of the digitalized Tamil text in our will be in these two coding. With Multilingual Domain Name System, in future, the Domain name and the URL can be typed in the Tamil language.

## References

1. Leong Kok Yong , Tan Tin Wee, Naa Govindasamy & Lee Teck Chee (June, 1996), Multiple Language Support over the World Wide Web, INET96 Conference Paper, Montreal.
2. Leong Kok Yong , Tan Tin Wee & Lee Teck Chee (March, 1997), Making your Web server render Unicode text for your client users, 10th International Unicode Conference, Mainz, Germany.
3. Bos, Bert (1996). Internationalization/Localization W3C: Non-Western Character Sets, Languages, and Writing Systems. http://www.w3.org/pub/WWW/International/.

4. Grimes, Barbara F., Editor. (1992). Ethnologue, Languages of the World. 12th Edition. Consulting Editors: Richard S. Pittman and Joseph E. Grimes. Summer Institute of Linguistics, Inc. Dallas, Texas.

5. Nicol, Gavin T. (1996). The Multilingual World Wide Web.http://www.ebt.com:8080/docs/multilingual-www.html.

6. Govindasamy, N. (1989). New Keyboard for Tamil Computer, by Naa Govindasamy, 7th International Conference of Tamil Studies Seminar Proceedings, Maritius, December 1989.

7. Govindasamy, N. (1994a). Computer and Tamil Teaching. 2nd International Conference of Tamil Language Teaching, Kuala Lumpur, June 1994.

8. Govindasamy, N. (1994b). Kanian Keyboard. Tamil and Computer Conference Proceedings, Anna University, Madras, India, August 1994.

# A GLYPH-BASED FONT ENCODING SCHEME TSCII AS A CANDIDATE FOR TAMIL COMPUTING

*K. Kalyanasundaram, Switzerland and M. Nedumaran, Singapore*
*Internet Working Group for Tools and Standards for Tamil Computing**

## Introduction

Dravidian Languages such as Tamil use non-Roman characters. Historically, transliterated form of writing Tamil text using Roman characters was the common practice, particularly amongst western scholars. Dediated softwares (text editors and word-processors) capable of rendering tamil scripts using built-in Tamil fonts made their debut in eighties and soon became popular particularly in Malaysia and Singapore region. These commercial softwares were rather expensive and so their usage was restricted largely to publishing houses. Free, self-standing font faces became available in the Internet in the early nineties and this resulted in a major explosion in the number of people who can handle materials directly in their personal computers. Though exact numbers are not yet available, it is likely that today at least a quarter of a million can handle tamil materials in tamil script on computers. This number is likely to double as another year passes by.

Last decade also saw phenomenal growth and establishment of Internet as one of the major modes of information interchange. Enormous amount of materials are being produced using computers using different operating systems and with different softwares. Information is also being exchanged through different exchange protocols (SMTP, NNTP, POP, HTTP, ...). Facile flow of information through different computers and protocols require in a mandatory way some standardized font encoding scheme.

A font encoding scheme is an explicitly written convention for handling of language script(s) by the computer systems. An encoding scheme can be glyph based where one uses a series of graphic characters (glyphs) stored at specific locations of a font to generate the script. All font (font faces) then use this same standardized format. An encoding scheme can also be character-based where one simply defines the basic elements required to define the entire alphabet and leave the details of rendering of the script to softwares. Unfortunately there has not been any single encoding scheme universally accepted for Tamil and used by everyone. This has resulted in mushrooming of hundreds of Tamil font faces (7-and 8-bit monolingaul and bilingual) in the Internet and near impossible situation of exchange of information between individuals. This paper proposes a glyph-based encoding TSCH as a possible candidate for fonts used in Tamil computing.

## TSCII Initiative

Two major developments were responsible for the present initiative. Almost three years ago, world-wide discussions started through two Email discussion lists "webmasters@tamil.net" and "tamilnet@tamilnews.org.sg". The participants with different background (software developers, typographers, linguists, academic scholars, users) are all interested to see an encoding standard defined soon. Earlier conference in this series TamilNet'97 held in Singapore (organized by the National Univ of Singapore) recognized the urgent need to define a standard and welcomed initiatives in this regard. The two mailing lists were soon merged into one. The present proposal TSCII is the outcome of discussions of an Internet Working Group for Development of Tools and Standards for Computing.

## Existing Tamil Standards

During this decade nearly all major representatives of hardware and software industry joined together and launched a global encoding initiative called UNICODE. Unicode is an ambitious character based encoding scheme to handle all world languages with specific segments allocated for each language including. While Unicode is nearly implemented for English and European Languages, it is not yet ready for India languages. Before Unicode, in the eighties, Govt. of India proposed an Indian Standard called ISCII to handle all Indian Languages under a single character-based encoding scheme. CDAC,

Pune developed softwares based on this ISCII standard and these are used in state and federal govt. establishments within India. For India languages, Unicode has adopted the character-based encoding schemes of ISCII. In spite of this parentage of Unicode with ISCII (as far as Indic languages are concerned), implementation of Unicode is quite different from that of ISCII. Neither the softwares nor text files/databases of one scheme applicable directly to the other.

Due to the difficulties in implementing ISCII in widely used information exchange protocols of Internet, CDAC recently proposed a "secondary layer standard" ISFOC. ISFOC is a glyph-based encoding standard with no direct inter-convertibility to the parent ISCII. The complexity of file transfers between these glyph-and character-based encoding standards takes away all benefits of cross-language advantages of this standard. So the situation is far from satisfactory. The complexity of these two schemes does not allow usage of any of the thousands of shrink-wrap softwares that are available for English language. ISCII scheme can be implemented only through dedicated hardware and/or softwares. Fortunately language is comparatively a simple language adequately handled by a glyph-based encoding scheme.

## Biggest Challenge of Today

The goal today is not to introduce one more font encoding to the arena. The biggest challenge facing us today is unification of hundreds of font encodings that are in use for today. Any standard font encoding scheme proposed today must allow facile migration of legacy documents, texts and fonts. Through its elaborate design goals, TSCII initiative attempts to address all these pressing requirements. The initiative recognizes that soon (possibly in a decade) Unicode will firmly establish itself as the world-standard for multi-lingual computing needs. Hence the initiative proposes a glyph-based encoding scheme as an "interim" option till Unicode is firmly established in all computer OS and softwares for the market.

## Design Goals of TSCII Initiative

The following are some of the key Design Goals of the proposed TSCII initiative:

i) the encoding SCHEME MUST BE IMPLEMENTABLE ON ALL COMMONLY USED COMPUTER PLATFORMS (Unix, Windows, Mac and others). New generationsof more powerful computers and associated softwares are being released every couple of years. The encoding standard should be such that it can be used in all computers released during the last decade (backward compatibility)

ii) the encoding SCHEME MUST BE OPEN. There will no need to get permission from anyone to implement the encoding standard in hardwares and softwares. No copyright restrictions of any kind. Nearly all of the International Standards for Information Interchange are OPEN standards.

iii) the encoding SCHEME WILL BE AN 8-BIT BILINGUAL (ROMAN/TAMIL). The widely used lower ASCII set (roman characters and punctuation marks) take its standard location (slots 0–127). Tamil glyphs occupy the upper ASCII berth (slots 128–255). This is based on the recognition that information exchange via widely used methods such as Email and Web is best assured if key information (tags) on the nature/content of the file are indicated using the usual lower ASCII set.

iv) the encoding SCHEME PROPOSED WILL BE GLYPH-BASED ONE WITH A UNIQUE COLLECTION OF GLYPHS to generate the entire alphabet. The encoding scheme should be such that there are no ambiguities in the interpretation of the resulting text (by search and sort engines for example), no redundancy nor repetition of old style of writing some alphabets (lai, Lai, Naa, naa and Raa). The encoding will allow text input as per the current practice of Tamils worldwide, without enforcing any language reforms. In view of the near absence of usage of numerals, their inclusion can be viewed as an exception. Two main reasons for this exception are: i) the need for the encoding scheme to handle very ancient manuscripts (Ftext archives) where these numerals were in use; ii) to allow join the main stream where many of the classical languages have their own numerals. Unicode does recognize this key fact in providing slots

for these in many language segments. The scheme will not include any idiosyncratic, personal, novel, rarely exchanged or private-use characters.

v) the encoding SCEME WILL BE UNIVERSAL IN SCOPE. While keeping grantha characters and numerals as part of the glyph choices, the encoding scheme is designed such that the basic glyphs are collected in widely used Latin–1 (8859–x) standards. This will ensure that the "pure Tamil" message gets through even in the poorest/bad local implementation scenarios. Supplementary characters such as granthas, numerals and rarely used alphabets (such as nju, ngu, njU, ngU) are to be placed in rows 8 and 9. By providing a couple of encoding slot positions as vacant, the scheme will allow software developers to use them as "escape routes" to bring in additional (special characters such as old-style Lai/lai/Naa/naa/Raa) characters if necessary.

vi) The encoding SCHEME HAS TO PROVIDE COMPATIBILITY to Unicode (and hence ISCII) scheme. Proposed inclusion of grantha characters and numerals is relevant in this context. Compatibility to Unicode will ensure easy migration of documents to and fro during and after the transition period.

vii) Efforts will be made to provide appropriate softwares that will ALLOW SMOOTH MIGRATION FROM HUNDREDS OF CURRENT FONT ENCODING SCHEMES. For conversion of legacy documents "Converters" will be provided that allow inter-conversion of files between widely used encoding schemes and proposed TSCII. The converters will also allow any user to quickly generate converter plug-ins for any custom encoding. For text/data input, several keyboard editors will be provided allowing input as per widely used keyboard layouts and existing font encodings. This way anyone can continue to do the input in nearly the same manner as before but all with the same encode-conformant fonts. Transition to TSCII thus will be smooth and rapid.

## Details of TSCII Encoding

**A possible 8-bit font encoding scheme for Tamil -v. 1.6**
**(Tamil standard Code for Information Interchange TSCII)**

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 (0-15) | | | | | | | | | | | | | | | | |
| 1 (16-31) | | | | | | | | | | | | | | | | |
| 2 (32-47) | | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| 3 (48-63) | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 4 (64-79) | @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 5 (80-95) | P | Q | R | S | T | U | V | W | X | Y | Z | [ | \ | ] | ^ | _ |
| 6 (96-111) | ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 7 (112-127) | p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | |
| 8 (128-143) | ௦ | ௧ | ஸ்ரீ | ஜ | ஷ | ஸ | ஹ | க்ஷ | ஜ் | ஷ் | ஸ் | ஹ் | க்ஷ் | ௨ | ௩ | ௪ |
| 9 (144-159) | ௫ | ‘ | ’ | “ | ” | ௬ | ௭ | ௮ | ௯ | ஙு | ஞு | ஙூ | ஞூ | ௰ | ௱ | ௲ |
| A (160-175) | | ா | ி | ீ | ு | ூ | ெ | ே | ை | © | ௗ | அ | ஆ | இ | ஈ | உ |
| B (176-191) | ஊ | எ | ஏ | ஐ | ஒ | ஓ | ஔ | ஃ | க | ங | ச | ஞ | ட | ண | த | ந |
| C (192-207) | ப | ம | ய | ர | ல | வ | ழ | ள | ற | ன | டி | டீ | கு | சு | டு | ணு |
| D (208-223) | து | நு | பு | மு | யு | ரு | லு | வு | ழு | ளு | று | னு | கூ | சூ | டூ | ணூ |
| E (224-239) | தூ | நூ | பூ | மூ | யூ | ரூ | லூ | வூ | ழூ | ளூ | றூ | னூ | க் | ங் | ச் | ஞ் |
| F (240-255) | ட் | ண் | த் | ந் | ப் | ம் | ய் | ர் | ல் | வ் | ழ் | ள் | ற் | ன் | | |

Figure 1 Presents in the form of a compact table glyph choices and their slot allocations (code positions) of the proposed 8-bit bilingual encodig.

It can be noted that the encoding scheme includes in addition to characters of tamil alphabet the following: single and double curly quotes and copyright sign at their code-positions of widely used Latin–1 scheme. Having these curly quotes allow ready usage of shrink-wrap software (for

word-processing, graphics etc.) that are available for use in English and European languages. Copyright sign (at its ANSI slot #169) is increasingly used in many of the Internet-based documents particularly in Web pages. Presence of this will avoid un-necessary switching to other Roman font faces. Two slots (254, 255) have been left vacant as "private use area" for software developers.

## TSCII Encoding and Desktop Publishing

Undoubtedly, the major use of the tamil fonts (Tamil computing) is in the word-processing application. Desktop publishing is becoming the widely used mode for printing even in professional publishing houses. During our 2-year long discussions, it was repeatedly pointed out that the glyph choices should be such that high quality printing required by professional publishing houses must be met adequately. Many of the tamil alphabets are complex forms graphically. It was pointed out that excessive use of kerning (as is the case with 7-bit fonts) renders delivery of high quality glyphs rather difficult. Since the number of slots available in upper-ASCII segment (#128–255) is much less than the required ones to allocate one slot for each of 240+tamil alphabets, choices have to be made on which of the Tamil alphabets are to be included in the native form, which are to generated using modifiers (several keystrokes in sequence). Choice of glyphs will determine the quality of the output and the slot allocations will determine the trouble-free performance across different computer platforms. It may also be pointed out the quality of any font face (outline definition of the glyphs) will largely determine the quality of the output. Most of the freely distributed Tamil fonts in the Internet perform poorly in this context. Whatever the encoding scheme the Tamilnadu Govt. adopts as the standard, the world of Tamil Computing will benefit enormously if the Govt. distributes free at least a couple of "high quality" font faces through the Internet.

Two factors guide the selection of glyphs. One is the FREQUENCY OF OCCURRENCE OF THE ALPHABETS in a typical Tamil text and second the STRUCTURAL COMPLEXITY OF THE TAMIL ALPHABET so that they can be generated nicely in on-screen display and in print with or without the add of kerning and other basic font handling techniques already available for over a decade in all computer platforms. It is not a good idea to go for an encoding scheme where 80% of the chosen glyphs occur for less than 30% of the actual text. Assuming that the quality of the glyphs in the font face are of exceptionally good quality, more the number of alphabets in native form, better will be the quality of the Tamil text. A good balance has to be made between frequency of occurrence and the structural complexity of some of the alphabets.

Fortunately several of the Tamil alphabets are written as a composite of two or three basic components (referred to as "modifiers"). e.g. aakara, ekara, Ekara, okara, Okara, aikara and aukara varisai alphabets. Along with the basic consonants (mei), it suffices to have a select collection of modifiers (aakara, ekara, Ekara, aikara, aukara) in the encoding scheme to generate all these compound (uyirmei) characters. There is no need to have these as unique glyphs in the encoding scheme. Similar logic can be applied to grantha series as well. It suffices to include the special ukara and uukara modifiers and can use the Tamil modifier glyphs for the rest. After an in-depth analysis of various options, it was decided to invoke modifiers for the "ikara" and "iikara" varisai alphabets and rest of the series are generated directly.

There have been several analyses of the frequency of occurrence of Tamil alphabets and they have been used earlier in determination of the keyboard layout. With the choice of glyphs discussed above (shown in figure 1), we have NEARLY 87.07% OF THE TAMIL CHARACTERRS ACCOMMODATED IN NATIVE FORM in the encoding scheme: meis with puLLis 28.85%; basic meis (akaram eRRiya meis) 23.50%; ukara varisai 11.88%; entire uyirs: 7.00 %; aakara varisai (with stand alone "aa" modifier) 6.39 %; aikara varisai (with stand alone "ai" modifier) 4.41%; eekara varisai (with stand alone "ee" modifier) 1.88%; ekara varisai (with stand alone "e" modifier) 1.44%; ti and tii 1.06%; uukara varisai 0.62% and aukara varisai (with e, au modifiers) 0.04%. It means that, nearly 87% of the Tamil characters are rendered as native ones without any kerning. Their quality will be dependent purely on the quality of the font face design.

Even in the ca. 13% generated via kerning (used mainly in the ikara and iikara varisai), majority of them can be generated in quite satisfactory way using kerning procedures. Kerning is a routine font handling technique now available in all of the common computer platforms/OS. As a right-end modifier, the ikara and iikara varisai uyirmeis can be rendered fairly precise on all platforms. So it

is likely that, using the proposed glyph encoding scheme, over 98% of the Tamil characters can be rendered easily on screen and in print without any loss of quality. Techniques such as pair-wise kerning can handle even the residuals adequately. Professional publishing houses with more stringent requirements on the glyph display invariably use more sophisticated printing equipment and high-end computer systems. Advanced font handling techniques such as glyph substitution (GSUB) through (or without) Open True Type fonts are already implemented at the OS level. Hence it should not be problem for these cases to use dedicated software where single form of these alphabets are stored elsewhere and brought in wherever they are needed.

## TSCII Encoding and Information Exchange

A second major area of application of the font encoding is information exchange through Email and WWW. We will discuss each of these one by one. With the emergence of 16-bit Unicode as the encoding for multi-lingualism, nearly all of the widely used computer operating systems now can handle correctly information at this 16-bit level. Computers released in this decade (target coverage of proosed TSCII) all can handle 8-bit encoded messages.

EMAIL: Nearly all of Email softwares (including those that are used in shell account access such as PINE) are 8-bit compliant. A routinely used communication protocol is MIME (also known as Quoted-printable or base 64 encoding). MIME was designed to allow email exchanges in all of the world languages without worrrying much into the details of the font encoding used. MIME simply sends the code positions of the characters (A0, B2, EA etc) and the user/client-software recodes the information as the local choice of font face and associated font encoding. Using MIME it is possible to exchange information across different computer platforms very reliably. During the testing of TSCII encoding, we have successfully used most of the commonly used Email softwares on all three computer platforms. We have already a handful of email discussion lists where TSCII-based tamil exchanges are taking place routinely and the participants use many different softwares running on Mac or Unix or Windows-based computers.

WEB: Current versions of both of the dominant Web browsers Internet Explorer a0nd Netscape are Unicode-intelligent. Using the "user-defined" case for the font-encoding, we have shown successfully that it is possible to present formatted tamil texts (TSCII based) in the form of Web pages. Users can read the tamil materials locally using their preferred web-browser and font face of his personal chocie. We have also shown successful demonstrations of information exchange of formatted tamil text materials via "portable document format (PDF)" on all three Mac/windows/Unix platforms. PDF format is increasingly becoming the preferred mode of distribution of formatted materials (e.g. catalogs and annual reports of business establishments).

## TSCII Encoding and Database Applications

Another major area of concern is on Database Applications. Any encoding scheme should allow facile search and sorting of stored/saved tamil information. The information searched could be a tamil text viewed on a word-processing application or large database in a business or governmental organization. Database handling can be considered in three stages; storing, sorting and searching. The data base could be directly as plain 8-bit text as per the TSCII encoding. Sorting of the 8-bit tscii data can be done through an intermediate layer where glyphs are substituted by de-coupled characters and use any of the standard sorting algorithms. A demo software "varisai" has been specifically written for this purpose and has been succesfully tested with TSCII based materials. Possible ambiguities between the glyphs and characters during "search" has been addressed in TSCII. One particular case worth mentioning is the presence of uyir letter "au" and also the corresponding "au-length" modifier in the TSCII encoding scheme. The presence of au-length modifier avoids the common use of mei letter "La" as the modifier character to represent the entire au-kara varisai uyirmei characters.

With the very likelihood of Unicode taking the place as the international encoding standard, alternate possibility would be in Unicode. An on-fly convertor associated with the application can convert tscii data → unicode for saving in the database and also render data back to the application via another unicode → tscii conversion. The intermediate layering can be transparent only to the application developers. Except for an encoding scheme that lists the entire 240+ alphabets in the required sorting sequence, usage of an intermediate layer in any glyph-based scheme is inevitable.

Unicode has already released some standardised sorting options as RFCs and there are already software developers working on developing softwares based on this option. Double-byte sorting has also been proposed as an otpion. Clearly there will be more than one way one can do the search and sorting.

## Concluding Remarks

The proposed glyph-based encoding is the outcome of nearly three years of discussions in a public forum accompanied by extensive field testing by a group of internet-linked volunteers group. It has been shown to be a very viable "interim" option for Tamil Computing. It has the support of a broad spectrum of Internet Tamil community. In a short span of two months since the present encoding scheme was adopted as the final form by the Internet Working Group, many Tamil commercial software developers have produced several TSCII-encoding based tools and softwares and agreed to distribute them FREE: Tamil font faces and keyboard editors for use in Windows, Mac platforms, text converters to go between TSCII and popular Tamil font faces and vice versa, Email software that allow exchanges directly in Tamil, On-fly convertors of web-pages, etc. A dedicated Web site for TSCII.

<http://www.tamil.net/tscii/>has also been formed to provide all the necessary technical assistance for quick implementation of the standard and to serve as "the site" where anyone can download above type of TSCII-based tools. Hence we strongly believe that the proposed standard is a very viable one, guaranteed to deliver the goods it promises. We sincerely hope that the Tamilnadu Government will give a fair hearing to this proposal and possibly adopt it for Tamil Computing as soon as possible.

---

* Internet Working Group is a voluntary group of individuals (software developers, typographers, linguists , academic scholars, users) who are deeply committed to develop tools and standards for Tamil Computing. Kalyanasundaram <kalyanasundaram@epfl.ch>and Nedumaran <muthu@murasu.com are the Co-Chairs of this Internet Initiative. Details on the TSCII encoding and various tools to use TSCII are available at the TSCII website http://www.tamil.net/tscii/>

# UNIFIED CODING SCHEME FOR TAMIL

*Dr. V Krishnamoorthy*
*Professor, School of Computer Science and Engineering*
*Ann University, Chennai 600 025, profvk@hotmail.com*

**Abstract**

*In this paper a coding scheme for Tamil is proposed. It is a monolingual glyph based font coding scheme. In contains a glyph based bilingual coding scheme and this in turn contains a character coding scheme. As such three different coding schemes, each one required for a different purpose can coexist, and conversion from one to the other is very simple.*

Chennai code Version 1.4

| | 32 0010 | 48 0011 | 64 0100 | 80 0101 | 96 0110 | 112 0111 | 128 1000 | 144 1001 | 160 1010 | 176 1011 | 192 1100 | 208 1101 | 224 1110 | 240 1111 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 0000 | | 0 | @ | ளி | | ளீ | | | | ளு | ள | [illegible] | ள் | [illegible] |
| 1 0001 | ! | 1 | கி | றி | கீ | றீ | | ' | கூ | கு | க | க் | ற் | அ |
| 2 0010 | " | 2 | ஙி | னி | ஙீ | னீ | | ' | ஙூ | ஙு | ங | ங் | ன் | ஆ |
| 3 0011 | # | 3 | சி | ஸி | சீ | ஸீ | ஸ | ‘‘ | சூ | சு | ச | ச் | ஸ் | இ |
| 4 0100 | $ | 4 | ஞி | ஷி | ஞீ | ஷீ | ஷ | ’’ | ஞூ | ஞு | ஞ | ஞ் | ஷ் | ஈ |
| 5 0101 | % | 5 | டி | ஜி | டீ | ஜீ | ஜ | 5 | டூ | டு | ட | ட் | ஜ் | உ |
| 6 0110 | & | 6 | ணி | ஹி | ணீ | ஹீ | ஹ | 6 | ணூ | ணு | ண | ண் | ஹ் | ஊ |
| 7 0111 | ' | 7 | தி | க்ஷி | தீ | க்ஷீ | க்ஷ | 7 | தூ | து | த | த் | க்ஷ் | எ |
| 8 1000 | ( | 8 | நி | 0 | நீ | 3 | ளூ | 8 | நூ | நு | ந | ந் | ஸ்ரீ | ஏ |
| 9 1001 | ) | 9 | பி | 1 | பீ | 4 | றூ | TM | பூ | பு | ப | ப் | ா | ஐ |
| 10 1010 | * | : | மி | 2 | மீ | 9 | னூ | | மூ | மு | ம | ம் | ெ | ஒ |
| 11 1011 | + | ; | யி | [ | யீ | { | டி | ீ | யூ | யு | ய | ய் | ே | ஓ |
| 12 1100 | , | < | ரி | \ | ரீ | \| | டீ | ி | ரூ | ரு | ர | ர் | ை | ஔ |
| 13 1101 | – | = | லி | ] | லீ | } | | | லூ | லு | ல | ல் | ற | ஃ |
| 14 1110 | . | > | வி | ^ | வீ | ~ | | | வூ | வு | வ | வ் | ன | ஞு |
| 15 1111 | / | ? | ழி | _ | ழீ | | | | ழூ | ழு | ழ | ழ் | று | |

## Introduction

In this paper a coding scheme for Tamil is proposed. This is a monolingual glyph based scheme. With very few modifications, it can be used as a bilingual glyph based scheme and also as a pure character coding scheme. Since one is seen as an extension of the other, they can coexist with easy interchange between them.

Chennai code Version 1.4 the following basic ideas are used.

1. It should be a monolingual glyph based font, so that implementation will be easy.
2. This code should make operations like sorting easy. For this, the character coding scheme as consonant + vowel is necessary. The conversion to and from the character codes to the glyphs should be fast. For this the letters corresponding to the same consonant should be strategically placed.
3. The character coding scheme, will be a part of the monolingual scheme, such that it can be proposed as a candidate for the Unicode scheme for Tamil.
4. The upper part of the coding should be used as a bilingual glyph coding scheme, with minimum modifications.
5. The interchange from one scheme to the other should be very easy, so that the transition to the Unicode will be smooth.

First we will see some of the points which lead to this coding scheme.

1. It is obvious that no glyph based font scheme can provide direct sorting sequence, with just 256 positions.
2. A monolingual font is essential for rendering the letters without distortion. When the necessity of bilingual fonts are questioned, it is mentioned that it will make things easier in places like multilingual viewing. An important point to note is, whether any newspaper or magazine publisher will be willing to print using a bilingual font? A standard should be usable by everyone.
3. Searching and sorting are the fundamental operations in any computing scenario. ( In general it is estimated that around 30% of the time, the computer is doing only searching and sorting. That is why so much research has gone into that leading to innumerable algorithms) The technology is changing so rapidly that Natural language processing in a language would be a must, if the language has to survive in the future. The code should help for this processing, and should not be a hurdle. When a code is proposed as Unicode we should keep in mind that this code will be used for a long time. It is also true that glyph based schemes are going to be there for some more years. As such the glyph based schemes will also be used heavily for purposes like spellchecking and in all other language processing applications, in the immediate future.

4. The best way to represent a letter is by two parts, the mei and the uir. The absence of mei and the absence of uir have to be indicated by special symbols.

The salient features of this scheme are explained below.

**As a Monolingual Font code -**

1. This is an ANSI compliant code. For this the 'non-printable positions' 127, 128, 129, 141, 142, 143, 144, 157, 158 and 160 are avoided. In order to avoid the problem posed by the 'intelligent quote marks' of some word-processors, the positions 145, 146, 147 and 148 are filled with quote marks.
2. Ten positions are given for the Tamil numerals. They are not in a sequence. But this is not necessary. The numerals for 10, 1000 and 1000 are not included in the scheme. If necessary they can be included in the Unicode version
3. The letters sa, sha, ja, ha, ksha, sri and trademark and bullet symbols have been included.
4. Note that bit manipulations are faster than table lookup. In English a lower case letter is changed into the corresponding capital letter by just changing one bit. This helps enormously in searching. The same technique can be used by us in a far more effective way.

   A consonant and its modified forms are kept such that they differ by a multiple of 16. This ensures that the conversion between a consonant and its modified forms can be done by manipulating the first four bits only, instead of table lookups. This will speed up the natural language processing while running as well as while programming, while using the glyph based font coding schemes. Except for a few letters, the remaining letters are in such symmetrical positions.
4. The positions 208 and 240 are filled by two special symbols. They will be treated just like the symbols ©. They may be related to the ekara, eekara symbols. But they will not be used as ekara or eekara modifiers. Note that no one will use the copyright symbol instead of the letter c. The same principle is used here. By this we have avoided the ambiguity These two symbols will be just seen when someone wants to see a text given in the bilingual format. These two positions will be filled by the actual ekara and eekara symbols in the bilingual coding format.
5. In the same spirit the two letters ti and tti are represented by two symbols.
6. In this monolingual font, all the Tamil letters can be created without breaking any of them. No kerning or negative offset is necessary.

**As a bilingual font code -**

The same scheme can be adopted as a bilingual font scheme also, with the following minimal modifications.

1. Fill the required 52 positions with the English alphabets. This removes the ikara, iikara sequence of Tamil letters.

2. The 6 Tamil numerals are also removed. For the sake of uniformity, do not use the remaining Tamil numerals also (positions 149 to 152).
3. The ekara eekarams are got by combining the symbols at positions 208 and 240 with the corresponding consonant symbols. The exceptions are ti and tii. They are placed in the positions occupied by the corresponding symbols in the monolingual scheme.
4. No other change is required.
5. A text saved in a bilingual font can also be seen using a monolingual font, without any change, since the two symbols at positions 208 and 220 in the monolingual font have a tinge of the ekara eekara symbols. With some simple conversions a monolingual font text can be seen using a bilingual font.

**As a character coding scheme -**

1. The positions for the uir, mei and the positions 208 and 240 are used to represent Tamil characters. The position 208 indicates the absence of the mei component. The position 240 indicates the absence of the uir part. Every character is represented as (mei part, uir part).
2. This character representation scheme will produce correct Tamil sorting. Also this is the most prefered scheme for language processing computations.
3. The positions 149 to 152 provided for the Tamil numbers 5 to 8 are used as they are. The positions 144 to 148 are filled with the Tamil numerals 0 to 4, and the position 154 is filled with the numeral 9. No other position is used.
4. This part can be proposed to form the Unicode. This character coding can use any database English sorting engine without any modification, to sort correctly in Tamil order.

**Advantages -**

1. This has the advantages of both the glyph based font schemes and the pure character based scheme.
2. This scheme is easily implementable as monolingual or bilingual fonts with the current technology. This avoids an unnecessary extra layer of software for rendering, as it is a glyph based scheme. Only the keyboard layer is required.
1. Since the monolingual, bilingual and character coding schemes share the common positions, transition from one to the other, and, saving to or retrieving from any scheme are easy. This will also minimize the problems involved in converting to the Unicode at a later date.
2. The natural language processing programs will be speeded up by this code, since it makes many fundamental computations just by bit manipulations. Also since the letters are not torn into pieces and thrown indiscriminately, but placed systematically, writing a program for natural language processing will be lot more easier.
3. If the 32 positions from 128 are lost due to bit stripping, only a few vookarams and some grandha letters will be lost. In the case of bilingual font, ti and tii also will be lost. The percentage of these is approximately 1 only. Hence not much information will be lost.

**Conclusion**

There are three types of persons dealing with Tamil code for computers. The first category is the font manufacturers for publishing work. Their requirement was just some placing of the characters. They did just that. The second category is people wanting to communicate in Tamil using computers. Their requirement is a common coding scheme. The third category is people working on natural language processing. Their requirement is not only a common coding scheme, but also an efficient scheme which will help in the creation of language products and will also speed up this processing.

There is a definite need for a monolingual font scheme for Tamil. A bilingual scheme can also coexist with this, only for some specific applications. Anyhow this need may be there only till the Unicode becomes wide spread. Interchanging Tamil letters by Internet should not be the only problem we should address now. With rapidly changing technology these problems may vanish quickly. But the problem is how we are going to do our natural language processing in the immediately coming years. Note that some of us are already working in this area. Our problems are real and are to be tackled today, now. The code proposed here is designed in that sprit.

# MULTIPLE ENCODING SYSTEMS FOR DIFFERENT COMPUTER APPLICATIONS

*Dr. M. Ganesan*
*Central Institute of Indian Languages*
*Manasagangotri, Mysore - 570 006*
*Ph: 0821-515558 (Off)*
*Ph: 0821-410740(Res)*
*e-mail: ciil@giasbgO1.vsnl.net.in*

## Introduction

Any language needs a standard encoding system for easy transportability of data across platforms (DOS, Windows, Unix, Mac, etc.) and application softwares. ASCII is one such standard for all the languages using Roman Script. There are many basic differences between these languages and Tamil. The writing system of the languages like English is basically a spelling based one whereas Tamil, for that matter any Indian language, is a syllabic language. Due to this difference and the complexity of Tamil scripts, there are problems in constituting a single encoding system. Again due to the lack of a standard code, many codes came to exist and created hindrances in data transportability and information interchange in Tamil.

## Standard Encoding System: Whether Glyph Based or Character Based

In information technology "*Characters* are abstract information elements in the domain of coding for data interchange" whereas "*glyphs* are abstract presentation elements in the domain of presentation processing" ....... Further "- A character conveys distinctions in the meaning or sound. A character has no intrinsic appearance. -A glyph conveys distinction in form or appearance. A glyph has no intrinsic meaning" (ISO/IEC Working Draft, 1996. page 3). It means that characters represent sound and meaning of a language whereas glyphs represent the writing of the language. Even historically speech preceeds the writing system of a language. "Writing.... is a late cultural invention and it comes later than articulacy in the history of individuals" and "writing is a device to record and a tool to represent speech" (Srivatsava and Gupta 1990:12-13). It is clear that the basic encoding system for information interchange must be a character based rather than a glyph based one.

## Glyph Based Encoding System

As far as Tamil is concerned, character based encoding system alone cannot meet the requirements of Tamil computing. The reasons are 1) not having one-to-one correspondence between character and glyph, 2) more number of graphic characters due to the syllabic nature of the language, 3) not having uniformity in consonant-vowel clustering and more importantly 4) limitation of technology at system level for handling the complexity of writing systems of Tamil like languages. For different applications like document preparation, database creation, e-mail, web page creation, browsing, and searching, etc. on popular platforms like Windows, at present, there is a need for glyph based encoding system.

## Disadvantage of Glyph Based Encoding System

Data on a glyph based encoding cannot be usable directly for a number of applications, listed under category III later in this paper. For any linguistic study on Tamil, characters have to be identified and recognized by the system, which is not easy with glyph based encoding system. Characters are represented with varying number of glyphs.

For example,

1) single glyph for two characters (phonemes)

    கு =k +u

2) two glyphs for single character

இ, ி =i

3) three glyphs for two characters

கொ =k +o, etc

That is, number of bytes per character is not uniform. So it is difficult to locate the character boundary. Sorting, which is a minimum requirement for word indexing, dictionary making, etc. is not possible with glyph encoding.

## Character Based Encoding System

As mentioned earlier, character encoding system is the basis for information interchange. Standard codes, ASCII, Unicode, ISCII, etc. are character based encoding system. The advantages are 1) sorting is simple, 2) easy data transportability across platforms and applications softwares and 3) Straight away usable for NLP related works.

## Short-comings of ISCII for Tamil Computing

ISCII code is a character based encoding system. It is a standard code common to all Indian Languages. But it cannot be taken as a standard for Tamil, for the following reasons:

i) The inherent 'a' in a consonant does not have a code. This is because of the considerations that க +˙ =க்; this derivation is unnatural. Because of the lack of code for 'a' matra, it is impossible to separate the vowel 'a' from a consonant. In Tamil, for example, 'a' is a morpheme denoting relative participle(RP) marker(peyarecca vikuti), etc

eg: படித்த = படி + த்த் + அ
verb + Past tense + RP marker

ii) Vowels and their corresponding matras are having two different codes, though they are phonetically same.

eg: நீராடு = நீர் + ஆடு

When segmenting 'a: Tu' from 'ni:r,'one will get ாடு which is first of all meaningless and while searching for ஆடு, ாடு will not be counted automatically, as they have different codes. Similar problems are there for all other vowels.

iii) ISCII follows the alphabetic order of Hindi, and therefore a number of characters are (ஜ, ற, ழ, ள, ண) dislocated from their regular alphabetic order of Tamil. So, sorting is not possible without an additional programming.

iv) Indian language numerals are not in the regular slots of Arabic numerals and therefore any computation, when it is needed, is not possible.

v) Each of the five plossive sounds (க, ச, ப, ட, த) has three more consonants as its varg- aspirated, voiced and voiced appirated- in all Indian languages except Tamil. In the case of Tamil, ISCII represents all the four consonants with single letter and thereby a consonant is encoded with four codes. Thus, unambigious representation of character is not possible with ISCII.

For the reasons mentioned above, ISCII cannot be adopted as a standard for Tamil. Similar is the case with Unicode, as it almost follows the pattern of ISCII. So, there is a need for a separate character based encoding as a standard for Tamil.

## Applications of Computers in Tamil Studies

It is mentioned in the proposal for TSCII[1](page 5) that "it is likely that over 90% of Tamil computing is in the form of simple wordprocessing of plain text." It is worth to mention here that Tamil texts have to be first inputted mostly through a word processor to have the data in machine readable form.

[1]1. A proposal for A Tamil Standard Code for Information Interchange(TSCII) submitted to Government of Tamilnadu by Internet Working Group for Tamil Standard Code(IWG–TSC)

A number of research Organisations and institutions have been working for the last one decade on various aspects of Tamil language. To mention a few, 1) the Central Institute of Indian Languages, Mysore has built a machine readable corpora to a size of above three million words for many Indian languages including Tamil; software tools for the processing of corpora have been developed and currently working on Lexical Resources development for 5 Indian languages including Tamil. 2)IIT, Kanpur and University of Hyderabad are collaborately working on a Machine Aided Translation between Tamil and Hindi. 3) C-DAC Pune is working on Spell Checker for Tamil. 4) Tamil University, Thanjavur has prepared a word index for Sangam literature. In most of these works Roman characters are used to represent Tamil, due to nonavailability of a standard character based encoding system. If an appropriate character encoding is not made available to the researchers, they are, further encouraged indirectly to use Roman characters. The greatest advantage of computer itself is that data are stored in electronic media and therefore any manupulations on the data can easily be done. Here the type of applications of computer on Tamil are broadly grouped under three categories.

**Category I:** Document preparation, Information exchange through Internet and e-mail, searching, CAI and CALL packages, Multimedia Titles, etc.

**Category II:** Calligraphic works - Desk Top Publishing in publishing houses, Newspapers, Magazines, Advertisement agencies, etc.

**Category III:** Sorting, Indexing, Information processing and retrieval, Corpus building including literary archives, NLP works which includes morphological analysis, syntactic parsers, semantic analysis, etc, Machine Aided Translation, (Electronic)Dictionary compilation, Language Standardisation, Style analysis,Text-to-speech(Speech Synthesis) and Speech-to-text (Speech recognition), etc.

The above classification is made on the basis of the requirement of different encoding systems for Tamil. They are

1. A Bilingual Glyph based encoding for category I
2. A Monolingual Glyph based encoding for category II
3. A Bilingual Character based encoding for categoy III

O. S. like windows has the feature of storing the text in 1) the format of the packages used 2) print-files and 3) ASCII format. It shows that data created can be stored in different formats depending upon the future use of the data. In case of Tamil, the same principle can be extended to the coding level, by considering the innate complexity of Tamil writing system and other facts mentioned earlier in this paper.

## Three Standards for Encoding of Tamil

1. Bilingual Glyph based encoding system: At present a number encoding systems are in use. I studied the characteristic features of four such bilingual encoding systems, produced by 1) ING-TSC(TSCII), 2) C-DAC, Pune, 3) Govindaswamy, Singapore and 4) Krishnamoorthy, Chennai. The chart given below compares the features of them.

| Sl. No. | Features | TSCII (1) | C-DAC (2) | Govindaswamy (3) | Krishnamoorthy (4) |
|---|---|---|---|---|---|
| 1. | Slots between 128 and 160 are left free | No | No | Only 3 characters in 150, 151, 157 | No |
| 2. | Whether all characters are represented or not | Yes | Yes | Yes | Yes |
| 3. | Whether character representation is unambiguos | expect | No | No | Yes |
| 4. | Tamil numerals included or not | Yes | No | No | No |
| 5. | Special symbols included or not | No | Yes | No | only |

The encodings is for bilingual use (i.e using Tamil and English together) for the applications listed under category I. So the features 4 and 5 are not seriously considered. All the four satisfy the feature given in 2. The features (1) and (3) are more crucial. If the slats between 128 and 160 are defined with Tamil characters, on terminals there is a chance for not getting them displayed. So, it is advisable that codes should be above 161. Generally the encoding system of Govindaswamy is better and only 3 characters, that too rarly used ones are defined in those locations. The other vital features to be considered is that the encoding is unambigious in generating graphic characters. Encoding system of Krishnamurthy is unambigious, if we take the glyphs defined only in Upper ASCII. To suit the requirements by keeping the constraints in view, a bilingual encoding system has been defined and proposed herewith. (Table 2.4.1). It satisfies all the three features which are necessary for the applications mentioned under category I.

## 2. A monolingual glyph based encoding

The purpose of this encoding is to meet the requirements of DTP works. If the encoding takes care of all the graphics characters, numerals, and special symbols with no or minimum of kerning, that will be a better suitable one for the purposes mentioned under category II. Though different monolingual encodings meet this requirements, the one proposed by Anbarasan M/s Applesoft has 3 additional features.

1. The fonts defined below 128, match with typewriter keyboard layout; 2) All the glyphs defined below 128 can make all graphics characters, and therefore, usable in the 7-bits systems like pagers, and 3) the alternant forms which was there earlier in use; Nai,'lai' and 'Lai' are also available in the system. Because of these merits, the monolingual encoding systems of Anbarasan can be taken as a standard.

## 3. Bilingual Character based encoding system

To my knowledge only Anbarasan, M/s AppleSoft has defined the bilingual character based encoding systems for Tamil, other than ISCII and Unicode. The character codes are defined such a way that 1) Tamil Characters occupy safely, above 193 as various standard packages often use some slots which are normally below 192, 2) Through bit manipulation Tamil characters will occupy only the slots of Roman Characters but not the symbols, 3) Tamil Numerals are assigned the codes of Arabic numerals, so that any computing can, even, be done with Tamil numerals. 4) Any addition of special symbols, need to be included can be done without disturbing the system. Further this encoding overcomes the shortings of ISCII and therefore it can be taken as a standard for character based encoding system.

## Conclusion

This paper establishes the need for three standards of encodings, till technology develops to handle languages like Tamil only with a character encodings. Provision have to be made for conversion from one encoding to another.

## References:


1. ISO/IEC Working Draft TRXXXXX, Information TEchnology- An Operational model for characters and glyphs 24 June 1996.
2. Srivastava. R.N. and R. S.Gupta, Diamentions of Applied Linguistics, Mysore: Central Institute of Indian Languages, 1990

**Table 2.4.1**
**Bilingual Glyph Based Encoding**

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | A | B | C | D | E | F |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | | | 0 | @ | P | ` | p | | | | ட | ஜு | ஞ் | ஜ் | ல |
| 1 | | | ! | 1 | A | Q | a | q | | | ா | கு | அ | ட் | ஷ | வ |
| 2 | | | " | 2 | B | R | b | r | | | 7 | சு | ஆ | ண் | ஹ் | ழ |
| 3 | | | # | 3 | C | S | c | s | | | ி | ! | இ | த் | க்ஷ | ள |
| 4 | | | $ | 4 | D | T | d | t | | | ı | ஞு | ஈ | ந் | க | ற |
| 5 | | | % | 5 | E | U | e | u | | | ீ | டு | உ | ப் | ங | ண |
| 6 | | | & | 6 | F | V | f | v | | | ρ | ஜூ | ஊ | ம் | ச | ஸ |
| 7 | | | ' | 7 | G | W | g | w | | | T | • | எ | ய் | ஞ | ஜ |
| 8 | | | ( | 8 | H | X | h | x | | | ூ | து | ஏ | ர் | ட | ஷ |
| 9 | | | ) | 9 | I | Y | i | y | | | © | நு | ஐ | ல் | ணை | ஹ |
| A | | | * | : | J | Z | j | z | | | ீ | மு | ஒ | வ் | த | க்ஷ |
| B | | | + | ; | K | [ | k | { | | | ெ | ரு | ஓ | ழ் | ந | கூ |
| C | | | , | < | L | \ | l | / | | | ே | ஜு | ஃ | ள் | ப | சூ |
| D | | | _ | = | M | ] | m | } | | | — | மூ | க் | ற் | ம | டூ |
| E | | | . | > | N | ^ | n | ~ | | | ை | ளு | ங் | ன் | ய | ஸ்ரீ |
| F | | | / | ? | O | _ | o | | | | டி | று | ச் | ஸ் | ர | |

# A PROPOSAL FOR TAMIL STANDARD CODE FOR INFORMATION INTERCHANGE

*N. Anbarasan*
*Applesoft, Bangalore - 560 010*
*Tel: 3386167, Telefax: 3357167*
*e-mail: aplesoft@vsnl.com, anbu.arasan@axcess.net.in*

## Introduction

Tamil is the only living classical language of the world. It creates interest among the researchers world wide. Various universities, Institutions and individuals around the world doing linguistic analysis, are involved in Tamil literature archive and analysis for various requirements. The efforts of these universities and institutions are funded by various Governments, semi-governments and non-government institutions, which involve crores of rupees. For the lack of a standard - which could meet their requirements, Tamil data portability is lost. The data of various projects are being archived in their own coding schemes, which is accessible only by those, who created such data bases. There is an invisible difficulty in exchanging the information and processing such information, by others - other than those who had created these data bases.

Computers have become an inevitable part of lifestyle of mankind. The Tamilnadu Government headed by Dr. Mu. Karunanidhi is having aspirations of making Tamilnadu a computer literate state. Computers are being introduced at school levels, it is the foundation to prepare trained manpower to meet the future requirements.

Amidst Computerisation at various levels in government offices, departments, undertakings, educational institutions etc. a vacuum felt in data portability of Tamil digital texts, be it a simple letter or a volume of literature. It is noteworthy mentioning that, wayback **during 1995 I came forward to offer my software christened Surabhi Professional ver 2.00 to the Government of Tamilnadu, free of cost to be distributed to the needy users by a centralised agency. However, a response is yet awaited.**

## Need for standard code

Tamil being a well structured and having well defined grammar even at word level, a need arises for everyone to convert the available data to a character level coding, which helps in various processing requirements. It can be mentioned that, the advanced technical development such as **spellchecker, translation, OCR etc., could happen only on the basis of Character coding,** not on the basis of glyph coding. **A character coding is a must for information processing and interchange.**

## Need for separate standard for Tamil

Tamil language is well structured with rich grammar. Tamil language is having different character order than any other Indian Languages, which can in no way be achieved along with any other language.

In principle the ISCII is designed based on Devanagari script (please note that ISCII is not phonetic or character coding scheme), which is having contrasting script structure than Tamil.

In unicode, all Indian languages are coded as separate standard, which will help to maintain the traditional identity of the languages. Tamil language is wrongly coded in the Unicode standard.

As far as ISCII is concerned, the revision of ISCII is ruled out in the near future. The way in which Tamil characters are coded in ISCII is a bottleneck in all the projects. With these backgrounds it becomes necessary to have a separate standard for Tamil.

Unicode will emerge as an international standard when the stand alone operating system like Windows 2000 comes supporting it or when the Windows NT becomes the standard desktop operating system. Microsoft has already developed Windows NT lab version with Tamil Unicode. If we could

comeout with a standard character code for Tamil, the Tamilnadu Government can appraise Microsoft in using the standard and push to Unicode.

## Need for Multiple coding Schemes

Any of the present OS can't handle Tamil character codes. These 0Ss are supporting fonts of different formats. However, there are certain application softwares which do not allow selection of fonts. One such application could be terminal emulation softwares. To meet the various requirements of the users, the proposal for Tamil standard consists of three coding systems, namely:

1. Character based coding
2. Glyph based, monolingual font coding
3. Glyph based, bilingual font coding

The ISCII is also having different coding system to meet different requirements. In any character based coding system, a font is a must for rendering of proper text.

## ISCII compatibility

The Government of India is spending crores of rupees on projects of National importance and on Indian Language enabling technologies. To make use of the benefits of such projects and for portability of data it is a must for any software developed for Tamil to support ISCII atleast in the form of export or import facility. Irrespective of the problems or shortcomings in the ISCII, it is a National standard for Indian Languages.

## 1. Character based coding

We are witnessmg a new horizon blooming for Tamil computing. There is a lot of projects initiated by the Govt of India, which is of national importance. These projects involve lot of processing on data, which is in Indian Languages. **It is evident from the National and International standards, the coding has to be based on characters and not on glyphs**. Following table shows Tamil character coding scheme.

| | 32 | 48 | 64 | 80 | 96 | 112 | 128 | 144 | 160 | 176 | 192 | 208 | 224 | 240 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 0 | @ | P | ` | p | | | | | | ஒ | | ள் |
| 1 | ! | 1 | A | Q | a | q | | | | | அ | [illegible] | க் | ற் |
| 2 | " | 2 | B | R | b | r | | | | | ஆ | டீ | ங் | ன் |
| 3 | # | 3 | C | S | c | s | | | | | இ | [illegible] | ச் | ஸ் |
| 4 | $ | 4 | D | T | d | t | | | | | ஈ | [illegible] | ஞ் | ஜ் |
| 5 | % | 5 | E | U | e | u | | | | | உ | [illegible] | ட் | ஷ் |
| 6 | & | 6 | F | V | f | v | | | | | ஊ | | ண் | ஹ் |
| 7 | ' | 7 | G | W | g | w | | | | | எ | | த் | |
| 8 | ( | 8 | H | X | h | x | | | | | ஏ | | ந் | |
| 9 | ) | 9 | I | Y | i | y | | | | | ஐ | | ப் | |
| 10 | * | : | J | Z | j | z | | | | | ஒ | | ம் | |
| 11 | + | ; | K | [ | k | { | | | | | ஓ | | ய் | |
| 12 | , | < | L | \ | l | \| | | | | | ஔ | | ர் | |
| 13 | - | = | M | ] | m | } | | | | | ஃ | | ல் | |
| 14 | . | > | N | ^ | n | ~ | | | | | | | வ் | |
| 15 | / | ? | O | - | o | | | | | | | | ழ் | |

## Data identification

In the present day computers, English is playing a vital role as every aspect of computer is designed using English. It becomes more than necessity to co-exist with English. This bilingual character based coding system is flexible in data identifsication. This is an important feature for data processing, communication, rendering etc.

### Information processing

Character coding provides Language analysis and processing facilities, which is the basic requirement for any further development of Tamil specific applications.

### Phonetic coding and archiving Tamil literature

As the coding is, in its basic phonemic level, it becomes handy to archive old and new literature alike. The display rendering system could take care of the script pattern to be used for the data. It means that, by using the phonetic coding system, the literature can be well preserved.

### Less space requirement

As the language is being coded in its basic form, the archiving of data will require less space.

### Phonetic spelling

The spelling of a word is the basic constituent alphabets (characters phonetic spelling is essential for grammar based processing such as spell checkers, dictionaries and OCRs. These phonetic spelling is also important in search operations.

### Portability

As the data is in its pure alphabetic code, the data can be shared amongst the various researchers, developers for their various requirements with converting the data.

### Word boundary

In applications, the minimum feature supported is search. It entirely depends on isolating words, this coding meets such requirement.

### Conjuncts

When two or more consonants occur together, these consonants join together to form a syllable, and take a different shape, which is very different from its constituent consonants. As these syllables are derived and can be converted to its costituent characters, are not coded.

### Abbreviations or signs

To increase the writing speed of Tamil script, certain abbreviations, were used. Theses abbreviations are difficult to decipher into character codes, without a lookup table. These abbreviations are coded as symbols. This is not exhaustive, and could be finalised to meet the requirement.

### Convertibility to 7-bit systems

There are installations of 7-bit UNIX systems, which are being used for linguistic analysis and processing. This coding system is designed in such a way that, with simple 'bit' operations or addition/subtraction operation, the data can be converted from 8-bit to 7-bit and vice versa.

### Usability on 7-bit systems

On conversion from 8-bit coding, the data is mapped on to the English alphabets, which is a must for processing.

### Numerals

Tamil is having its own distinct shape for numbers, as found in any other Indian Languages. As per constitution of India, it is mandatory to use shapes of international numbers. As for computer processing, the shapes are not important, but the coded value. Therefore, Tamil numbers are retained in place of international numerals.

### Punctuation marks

The punctuation marks used in English is retained, as they also have meaning related to computers.

## Converging Technologies

Computer being the nodal centre for any digital communication, a phonetic coding is a must for bridging the gap in the communication protocol and their implementation platform.

## 2. Glyph based, monolingual font coding

As the present OSs are font based system, a simple font based coding will help the Tamil users to avail the OSs feature to have Tamil on their computers. This kind of simple minded coding will help spreading usage of Tamil on computer and help in exchange of data.

**This Glyph based font has been used in Rural Reach Project of Infosys, to create computer awareness amongst the rural primary school children** as part of their social service. This programme was carried out for Tamil and Kannada.

The glyph based font coding scheme based on basic shapes of Tamil letters will help the thousands of school children, and the Tamil literate masses. **This is the only way which can lead to Digital Tamil revolution, Governance in Tamil and Tamil communication on computers.**

The glyph based coding system can't meet any of the processing requirements because of the non-linear writing method of the Tamil language. In the absence of any processing based solutioons in the glyph encoding, the proposed monolingual glyph coding meets certain inevitable requirements. The glyph coding is shown in the following table:

| | 32 | 48 | 64 | 80 | 96 | 112 | 128 | 144 | 160 | 176 | 192 | 208 | 224 | 240 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | | 0 | @ | • | ங | ா | | | | டு | ஜ | ளீ | ன் | ழு |
| 1 | ஸ | 1 | ~ | ஈ | ய | ஜு | | ' | ஸி | ஷ் | யீ | ழீ | யி | ழி |
| 2 | ' | 2 | இ | க | ெ | ச | | ' | ஜீ | கி | யு | ீ | ய் | ழ் |
| 3 | ஃ | 3 | ண | ஞு | உ | ள | ஓ | ' | ஜீ | கி | யூ | பீ | ட் | பி |
| 4 | ஐ | 4 | ஜு | கூ | ள | வ | க்ஷ | " | ஜி | ஷ | றா | பு | ற் | ப் |
| 5 | , | 5 | நு | ரு | ந | ர | ஸ்ரீ | | வி | ங் | நீ | பூ | றி | வி |
| 6 | ஆ | 6 | கு | ஏ | க | எ | ! | – | வ் | ஙி | ளீ | ஞ் | ளி | ஞி |
| 7 | • | 7 | ழு | று | ப | ற | & | – | வீ | · | தீ | ஜு | தி | ஞ் |
| 8 | ( | 8 | ழ | ஓ | ா | ஒ | ^ | " | ஜ் | ஙீ | ீ | வு | த் | நு |
| 9 | ) | 9 | ஐ | ஜு | ை | ை | % | ஓ | © | ஔ | நீ | ஜூ | தி | ஷ் |
| 10 | • | : | து | ூ | த | ண | $ | மீ | க் | ஷி | நூ | ரீ | த் | ரி |
| 11 | + | ; | மு | ] | ம | { | ரு | ஔ | « | » | ஜூ | ரு | ன | ர் |
| 12 | . | < | டு | ˉ | ட | , | | ரூ | ஸ் | ஔ | லீ | லீ | ன் | லி |
| 13 | - | = | ஆ | ] | அ | } | | | - | ஷி | மீ | ஜூ | மி | ல் |
| 14 | - | > | கூ | ு | ே | கூ | | | ● | × | மூ | லீ | ம் | ரூ |
| 15 | / | ? | டீ | T | டி | | | ீ | ச் | \ | ரூ | ளி | ஜீ | ஜூ |

## 3. Glyph based Bi-lingual font encoding

For the obvious reasons, the OS doesn't support character based coding. To have a standard for complete requirement, a bilingual font is also being suggested. To keep rolling till such facility is available, an interim bilingual coding is designed to meet the requirement of programming, internet surfing and terminal communications. Except terminal communications, which is not possible without a bilingual font, the other requirements such as programming and internet surfing (specific to Tamil) could be achieved even with a monolingual font

Already there are atleast two proposals for bilingual font encoding, one is from Naa. Govindasamy, National University of Singapore and the other is from IWG-presented by Muthu. These two coding systems are following different approaches in coding. It could be summarised as:

1. To use all upper ASCII-as followed in the IWG proposal.
2. To use only above 160-as followed in the Tamilnet - Kanian coding.

On evaluation of the merits of these coding, a bilingual glyph coding be finalised.

## Conclusion

It is nevertheless to mention that each and every soul who has tried or trying to offer some valuable service to meet a fascinating standard for Tamil computing deserve appreciations and the services need encouragement.

The proposals or some services for reaching the set out destination and leading the Tamil computing with the trend setting technology into the next millennium, may come from any direction, they would be directed only towards a single and common goal of sophistication of Tamil computing in the IT revolution.

I am submitting this proposal just to enlighten certain points on standardisation of Tamil. I do hope, this proposal is worth considering for the noble cause of 'Digital Tamil'.

# TAMIL, THAMIZHEE AND NADODI

*Manoj Annadurai*
*mano@chennaikavigal.com*
*Benjamin Arulkumar*
*benjamin@chennaikavigal.com*

**Abstract**

*Statistics show that less than 2% of Tamil Nadu's population use computers. That too it is reserved to the privileged, elite class, that can comprehend English. To access the other 98% it is imperative that the software be available in Tamil. Software with even the user interfaces (UI) like Menus, Command Buttons, List Boxes in Tamil face certain daunting problems. This paper discusses the various problems thus encountered, proposes certain standards, suggests protocols that can be adopted. These suggestions are based on real-world user inputs.*

## Introduction:

Computing FOR Tamil. Computing IN Tamil. It is not playing with words. The difference is in the design. The distinction is between the ideology and technology.

With the positive thrust given by the Tamil Nadu Govt. towards IT (Information Technology) in general and Tamil computing in particular a possible real-world scenario would be that of mushrooming internet booths in and around the state. But do we have the software? Is there a Tamil Browser? A Tamil Spread Sheet program? A Tamil Accounting package? Or would we force a person to learn English to use his computers? As we bring about software IN Tamil, What are the standards to be set? Protocols to be adopted? Issues to be resolved? Our experience in developing high-end word-processors, spread sheets, 4GL Databases, Browsers (where the user can even type in URL's in Tamil)... gave us some unique and meaningful insights. This propelled us to adopt some schemes and protocols, that we would like to share now.

## Issue 1 : Save As HTML

Selection of Tamil words for user interfaces (UI). Though we have developed cutting edge technology for Tamil computing, naming them in Tamil has always been a problem. Now, how do we give "Save As HTML" in Tamil? Would it be **ஹச்.டி.எம்.எல்.லில் சேமி** (or) **இணைய மொழியில் சேமி ?**. While naming UI in Tamil we can follow, I.B.M.'s CUA/SAA ("Common user architecture/ System Application Architecture") advanced user guide. While we translate a word we have to bear in mind, that the translated word must be shorter than, if not equal to the English word. Also some times we can avoid a direct translation. For example, a direct translation of "Exit" would be "**வெளியேறு**". However, in our applications we use "**நன்றி**". Again for the menu item-"Tool Box", the word "அஞ்சறை" is used instead of "கருவிப்பெட்டி". This method- by Prof. Narayanan- of bringing the essence of the English word, into the beautiful spectrum that Tamil offers; such that even native Tamil speakers can easily visualize the objects and actions in a windowing environment, as described in IBM's CUA/SAA advanced user guide, is highly commendable. Some words that emulate the above model are: "**மின்னல்**" an acronym for "**மின். அஞ்சல்**" (E-Mail) and *தமீழ்* (தரவு மீட்பு மொழி) a Tamil SQL. But the story has just started. We have developed a technology that can bring scores of statistical, financial, trigonometric

function on to a Excel compatible, Tamil spread sheet. However, we are stumped while we attempt to translate functions like Z-Test (which returns the two-tailed p-Value of a Z-Test) . Even the book entitled "கணிப்பொறி கலைச்சொல் அகராதி" published by Anna University, though a laudable effort, does not contain these words. Therefore, we humbly request the government to setup a committee, a think tank, exclusive for this. It can have Tamil scholars like Dr. Elavarasu, Dr. Porko, Tamil linguists like Prof. Narayanan academicians like Dr. Ananthakrishnan, Dr. Ponnavaiko, Dr. Krishnamoorthy, and people in the Tamil software industry. (For screen shots refer Annexure)

## Issue 2 Why? What? Where?

**Why** we need a common font has been discussed endlessly and is beyond the scope of this paper. Therefore, We'll discuss **what** a common font mapping should achieve. To do so, we need to know **where** a font is used. Primarily a Tamil font was used in D.T.P, word-processing and information interchange. Now-a-days, we use it in database manipulations, spread sheets, creating web pages, browsing, accounting and for other purposes as well. As of now, for Tamil there are two main contenders (i) a monolingual scheme and (ii) bilingual scheme.

It is a well known fact that, it is the publishing industry that spreads Tamil, among others. Now, what kind of a font does a publishing industry use? Mr. Chellappan of Palaniappa Brothers, a giant publishing house (Publishers of the famous கோனார் உரை), printing more than a million Tamil books annually has this to say: " Bilingual fonts has absolutely no space in print media, because of lack of quality ...".

Leading dailies like "Dina Malar", "Dina Mani" use monolingual fonts. In all probability the book in which this would be published will be using a monolingual font.

Facts show that information interchange via e-mail can be done using a mono lingual font. The very few systems in which this might not be possible, are legacy. Current software standards and new versions, are in favour of monolingual fonts even for information interchange through e-mail.

It is also pertinent to know that **ALL** major software developer's in Tamil Nadu use a monolingual scheme. More importantly even the fonts used by the Tamil Nadu government, are monolingual in nature.

If at all we have to compromise, Chennai code 1.4 by Dr.Krishnamoorthy can be adopted. It matches all the TSCII 1.6 characteristics and simultaneously acts as a monolingual font. Nevertheless, a compromise is a compromise. Taking into account a lot of parameters, all the needs for applications developed today, and also for Tamil computing tomorrow, our preferred choice for character placements in a standardized Tamil font is as shown in fig.1. This font encoding scheme is much like Chennai code 1.1.

## Issue 3: முதலமைச்சர்@தமிழ்நாடு.அரசு

Elamaran is very excited on hearing about the internet, he wants to thank his beloved Chief Minister for bringing an internet booth in his village. In order to type the URLs and e-mail Id's in Tamil we propose two models.(Refer Annexure)

1. A server side implementation and / or

2. A client side implementation.

### Server side implementation

All Tamil based web sites will have to register their Tamil domain name with the ISP. Thus DNS information is stored in Tamil ensuring that the browser does not require a translation, when a user types in a URL. Using a similar concept with a considerable investment of time and money will ensure all web sites will have Indian Language translation of domain names in each of the ISP locations. The DNS entries are translated by the ISP's DNS server. Option can be given for server side translation / transliteration . In the case of E-Mail addresses using a server side technology, IDs cannot be stored, in Tamil. In this case, client side implementation is necessary, for the time being.

### Client side implementation

For a client side implementation we need an airtight Tamil to English transliteration scheme. The problem is, one is a phonetic language while the other is not. In English we have too many occurrences of different words spelt in the same way but pronounced differently. "Ch" is pronounced as "K" in Chemistry whereas it is pronounced as "Sh" in "such". Many a times, even the converse holds true: "mi" in "micro" and "my" as in "myself" are pronounced the same but spelt differently. A Transliteration scheme that takes into account the above facts, is currently being developed by Prof. Narayanan, an associate editor of Crea Dictionary. However, this cannot happen over night. Hence, we have implemented an interim scheme called Narayanan scheme 1, that approaches the problem in a radically different perspective. Each character in the English alphabet is assigned a Tamil equivalent as shown in fig.2.

Thus www.elcot.com would be **உஉஉ.எல்சஒட்.சஒம்.** The limitations in the interim scheme are obvious. To develop the perfect scheme here again, we request the state government to setup a committee along with the pioneers of the scheme. Such a transliteration scheme will not only be used in web browsers or e-mail clients, but can be put to use for many advanced applications like text-to-speech engines, speech-to-text processors and many other softwares. This scheme can translate decades of research and development into award-winning application in Tamil,overnight.

As the sun sinks into the horizon, at the twilight zone, when the evening turns into darkness; on 8th of February 1999, Tamil computing in it's truest sense would dawn. We would have a key board layout that could never ever be better. A font map that would put Tamil ahead of other languages. At this space-time continium a great responsibility is thrust upon our shoulders. The standards we choose now, would affect the usage of Tamil computing tomorrow. We now know the kind of food Sir Issac Newton had centuries ago, the exact words that Einstein spoke decades back... for Time records. And if Time records, Time will not delete.

## Cognizance:

We would like to thank **Dr. Ananthakrishnan,** who encouraged us to write this paper. He is our constant source of inspiration . We also like to thank **Mr. MuthuNedumaran** and **Mr. Mani.M.Mannivannan** for many fruitful dialogues. **Mr. Mannivannan** in particular for a lot of useful suggestions.

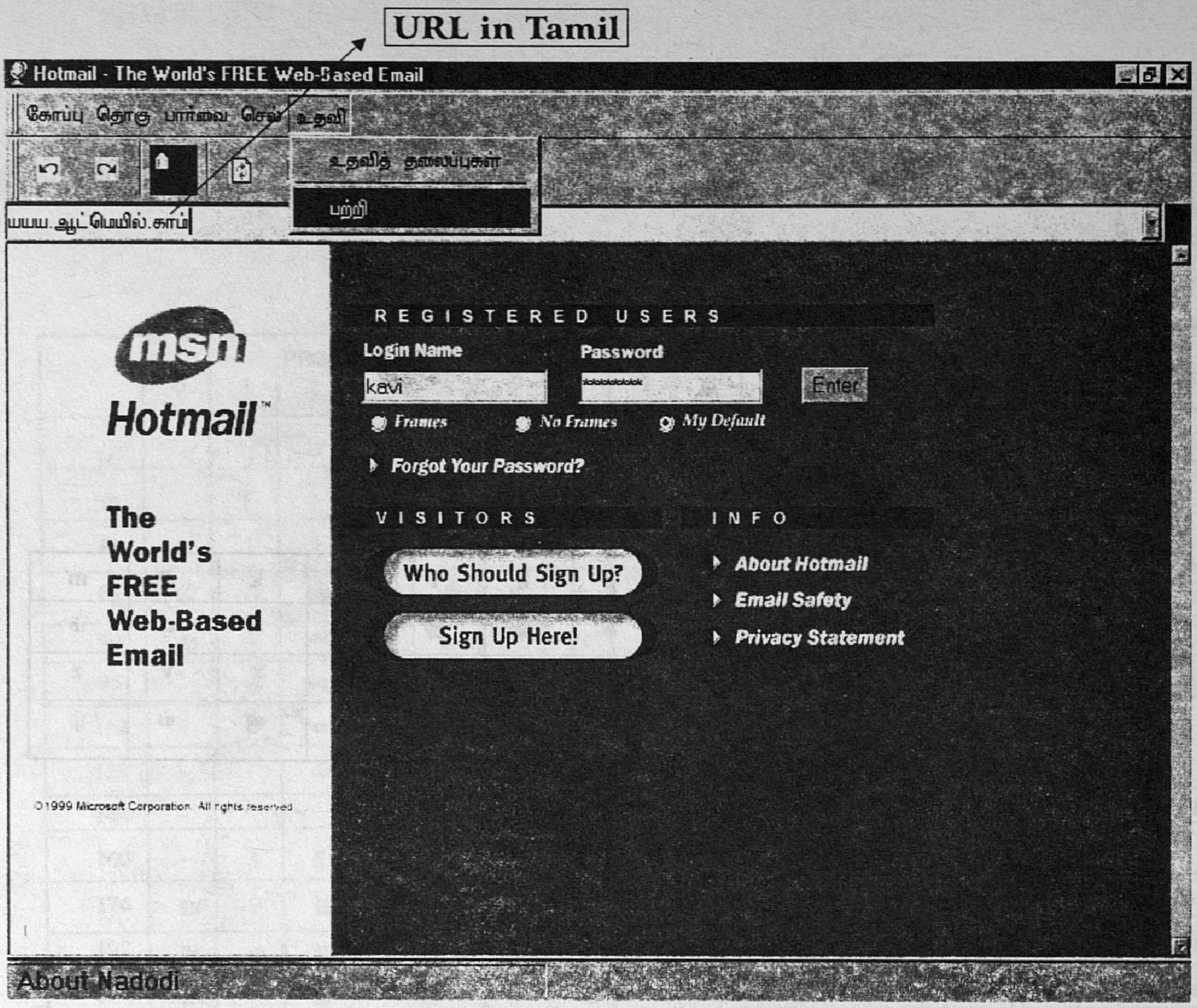
## PROPOSED STANDARD FOR COMMON FONT MAPPING (Fig.1)

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 16 | | | | | | | | | | | | | | | | |
| 32 | | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| 48 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 64 | @ | ஃ | அ | ஆ | இ | ஈ | உ | ஊ | எ | ஏ | ஐ | ஒ | ஓ | க | ங | ச |
| 80 | ஞ | ட | ண | த | ந | ப | ம | ய | ர | ல | வ | [ | \ | ] | ^ | _ |
| 96 | ` | ழ | ள | ற | ன | க் | ங் | ச் | ஞ் | ட் | ண் | த் | ந் | ப் | ம் | ய் |
| 112 | ர் | ல் | வ் | ழ் | ள் | ற் | ன் | கி | ஙி | சி | ஞி | { | \| | } | டி | - |
| 128 | - | - | ணி | தி | நி | பி | மி | யி | ரி | லி | வி | ழி | ளி | - | - | - |
| 144 | - | ' | ' | " | " | றி | னி | ா | கீ | ஙீ | சீ | ஞீ | டீ | - | - | ணீ |
| 160 | - | தீ | நீ | பீ | மீ | யீ | ரீ | லீ | வீ | ழீ | ளீ | றீ | னீ | கு | ஙு | சு |
| 176 | ஞு | டு | ணு | து | நு | பு | மு | யு | ரு | லு | வு | ழு | ளு | று | னு | கூ |
| 192 | ஙூ | சூ | ஞூ | டூ | ணூ | தூ | நூ | பூ | மூ | யூ | ரூ | லூ | வூ | ழூ | ளூ | றூ |
| 208 | னூ | ஸ | ஜ | ஷ | ஹ | க்ஷ | ஸ்ரீ | ஸ் | ஜ் | ஷ் | ஹ் | க்ஷ் | ஸி | ஜி | ஷி | ஹி |
| 224 | க்ஷி | ஸீ | ஜீ | ஷீ | ஹீ | க்ஷீ | 0 | கு | உ | ங | ச | ரு | கூ | எ | அ | கூ |
| 240 | ா | ீ | ு | ூ | ெ | ே | ை | _ | ~ | ஔ | | | | | | |
| 256 | | | | | | | | | | | | | | | | |

**Narayanan Scheme. (Interim) (Fig.2)**

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| a | b | c | d | e | f | g | h | i | j | k | l | m |
| அ | ப | ச | ட | எ | ஃ | க | ஹ் | இ | ஞ் | க் | ல் | ம் |
| n | o | p | q | r | s | t | u | v | w | x | y | z |
| ன் | ஓ | ப் | ங் | ர் | ச் | ட் | ஷ | வ் | உ | ஸ் | ய் | ழ் |

(Annexure)




# About the Authors:

Benjamin Arulkumar is a specialist in making different computers, running different operating system, talk with each other. He works as a Networking consultant, for some big firms in Chennai. He is instrumental in porting a lot of stand-alone Tamil applications to the Networked World.

He is now working in a project for a major firm in South India where, when a user logs on to the network for the first time a beautiful Thirukkural would be displayed on the screen. The user is also given options for audio , translation etc., This is being implemented enterprise wide.

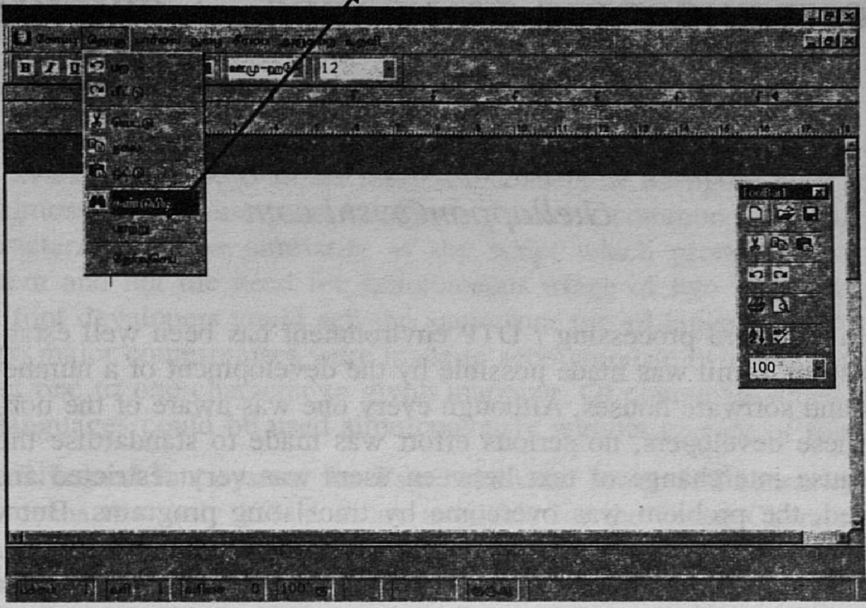
Manoj Annadurai is an avowed Tamil computing enthusiast. He with his colleagues started a firm called Chennaikavigal, which pioneered Tamil computing. Some of their software include:-

**Padhami**-The first Win95 based wordprocessor in Tamil.

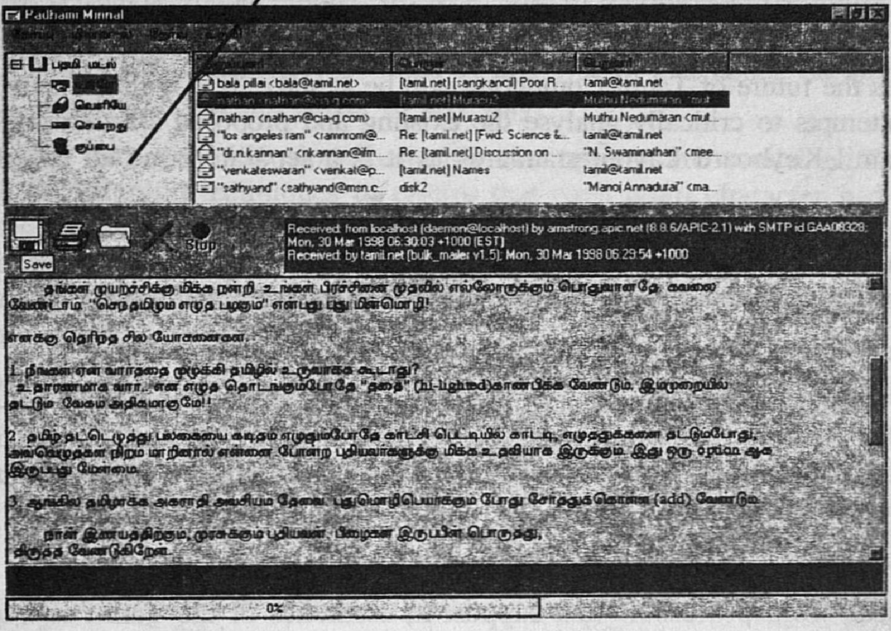
**Kuralamudhu**-The first multimedia encyclopedia on Thirukkural in a CD-ROM.

Currently they have developed a state-of-the-art Tamil word processor with spell checking and sandhi checking . They have a browser which accepts URL's in Tamil and an E-Mail client in Tamil.

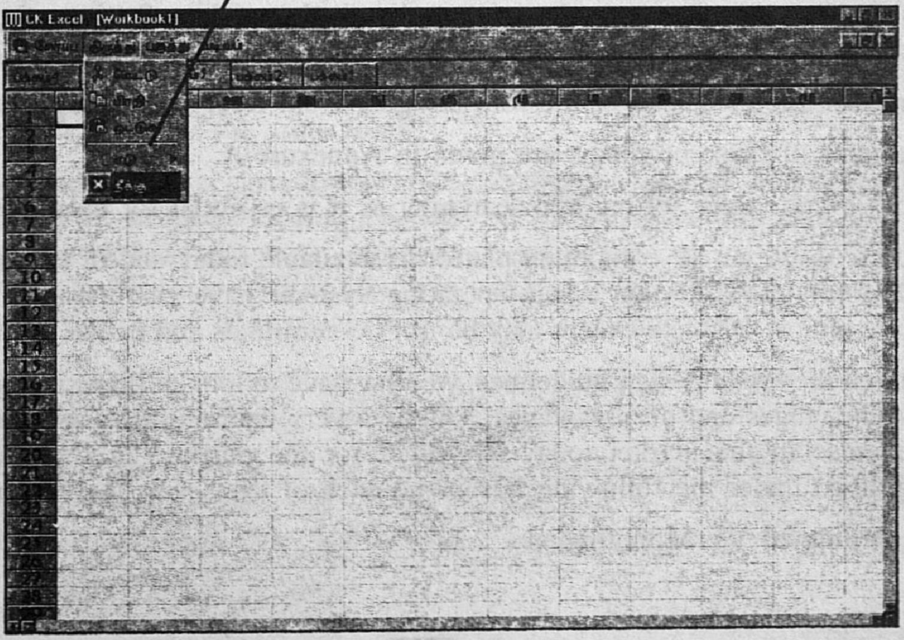
Padhami 2 - Menu in Tamil



Minnal - The first e-mail client in Tamil




Now even spread sheets in Tamil



# TAMIL FONT ENCODING STANDARDS – A CRITICAL STUDY

*P.Chellappan,*
*Partner, Pallaniyappa Brothers, Chennai,*
*chellappan@vsnl.com*

## Synopsis

The use of tamil in the word processing / DTP environment has been well established for nearly a decade now. The use of Tamil was made possible by the development of a number of fonts by many enthusiastic tamils and software houses. Although every one was aware of the non-standard encoding schemes used by these developers, no serious effort was made to standardise them as it was not a major concern because interchange of text between users was very restricted and in the few cases were it was required, the problem was overcome by translating programs. But with the advent of Internet Computing in the 90s people started communicating with each other and suddenly everyone started feeling the urgency for standardisation. Fortunately or unfortunately we are in the process of standardising at a time when the world is moving towards a global standard (UNICODE) for information interchange. This is a transition stage and one has not only to think of the present but also the future while evolving a standard. Another major factor to be considered during this process is the fact that in the near future the use of Tamil in computers will not only be restricted to mere word processing or DTP but also database operations. Hence if due and proper consideration is not given to these major issues the future of Tamil Computing will be at stake. It is with this serious concern in mind that this paper attempts to critically analyse the existing and proposed standards. This paper does not deal with the Tamil Keyboard Layout standards as it is a separate issue all together.

## Introduction

Before proceeding to analyse the available and proposed standards, one has to decide on the criteria or design goals these standards have to meet. Any standard must necessarily satisfy the following conditions:

1. It must be capable of accommodating all the Tamil and grantha characters and their variants.
2. It should enable unique character to glyph mapping as far as possible because the quality of the printed text is as important as text interchange. Any attempt to create character representation by combining glyphs through keming will greatly affect the quality of the printed output because:
   a) Not all word processors support kerning.
   b) In those word processors that support it, aesthetics of the Tamil character will be lost.
   c) Softwares with only basic justification capabilities pull the combining glyphs apart. This is not a desirable feature in the print media. And even in softwares where you can prevent addition of intercharacter spacing during justification, the look of the output will suffer.
   d) It inhibits the creativity of the font designer from freely designing character shapes.

   Examples of these problems are given in Annexure-A.
3. It should facilitate easy sorting and searching as it is essential for database applications.
4. It should be usable on all computing platforms available today and of course without doubt on future platforms too. The text based Operating Systems have given way to the graphics based OS and are almost obsolete except may be in POS terminals and hence need not be considered.

Having stated the broad design guidelines, we now have to look at the existing encoding schemes in terms of its advantages and disadvantages. This is a very daunting task as there are "N" number of different schemes in the market today. However for the purpose of this study these different schemes are grouped under the following heads and each of the heads analysed separately :

1. Bilingual Vs. Monolingual

2. 128 Characters Vs. 256 Characters

3. Single Byte Vs. Double Byte

**Bilingual Vs. Monolingual :** In Bilingual fonts Character Sets of two different languages co-exist and they share the available 256 slots in the font. The concept of Bilingual fonts came into existence in Europe where almost all the European languages shared a common script and varied only in a few accented characters. It is the similarity of the script which prompted the emergence of the Bilingual font system and not the need for simultaneous usage of two languages. A Bilingual font system meant that font developers could sell the same font for all languages. At the same time and more importantly no major compromises were made to accommodate two languages in the same font. There was a perfect one to one character to glyph mapping. Of course an added advantage was that several European languages could be used simultaneously without change of font.

The use of a Bilingual font system for Tamil with the lower 128 character set being used by the Roman Script means that only 128 slots are available for the Tamil Character set and even out of these slots a few are unavailable because of some OS restrictions. In the Multilingual font system as envisaged by the ISCII standard too only 128 slots are available.

Although the Tamil language has 256+ characters which are unique, the design of these individual characters fortunately allow them to be constructed by adjacent placement of two or more component or glyphs. By using this advantage we were able to accomodate the entire Tamil character set in a 256 character system and at the same time did not compromise on the individual characteristic of each character. Kerning was completely avoided in this system and hence the print quality of the text and the display quality too did not suffer in any way.

If we have to live with only 128 slots then only two options are available. The first is to use sophisticated Glyph Substitution techniques to ensure that each tamil character had a unique glyph and the second is to compromise on the quality of the printed text. The first altemative is just evolving and is currently not available on low end software packages. Hence only those users who can afford these expensive packages can benefit form it. The second altemative is really not an altemative and is only a poor man's choice to say the least.

Having said that, the pertinent question now is "Why Compromise at all?". Why should English be used along with Tamil in the first place? Is it to satisfy the needs of a few who would like to use English along with Tamil without changing the font? In the few cases where arguments are forwarded for the necessity of co-existence of English and Tamil can we not find a work around? Is it really impossible to use Tamil on the Intemet without providing for the English script in the Tamil font? Is it because of convenience that we choose to have a bilingual font? Is it that this is the only solution which will ensure a trouble free use in database applications? Should the quality of the Printed Output suffer because of this unnecessary compromise? By using monolingual fonts have not English-Tamil Dictionaries been printed by using available off the shelf DTP packages? Have not ordinary users printed great looking letters and articles containing Tamil and English text by using monolingual fonts even on low end word processors? Should the Publishers and Printers, who produce millions of books and magazines, go in for expensive software just to satisfy the needs of a few who need English along with Tamil in the same font?

The above questions which repeatedly come to our mind cannot be simply ignored or wished away. The argument that by using a bilingual font like TSCII vl.6 one can print a text that has nearly 87.07% of its characters looking correct is simply not acceptable. Does it mean that the remaining 13% of the characters should be overlooked because of the compromise? Is it acceptable to us if 87% of an image in a photograph is in good focus and the remaining not?

Well, compromises can be made and should be made if there is no other altemative and if the future of Tamil is at stake. But till then let us not talk or even think about compromises.

One final thought on this subject. If the Tamil scholars agree to simplify the tamil script and form words just by placing consonants and vowels like the Roman Script then of course it will be possible to even have a multilingual font without any compromises instead of a bilingual one !!

**128 Characters Vs. 256 Characters :** Most of the above arguments in favour of monolingual fonts apply here too. Suffice it to say that only a 256 character based monolingual font will be capable of meeting the needs of the Tamil Text Processing enviromnent.

As far as the database applications go both the monolingual and bilingual font systems based on a 128 character set will pose varying degree of problems in areas for sorting and searching which is so vital to these applications. But several attempts have been made and it is definitely possible to arrive at a solution once we decide on a standard character encoding scheme. This may not be a big stumbling block.

**Single Byte Vs. Double Byte :** The double byte system has been widely used by the Chinese, Japanese and Koreans. This has helped them to accommodate the several thousand characters in their script. The far eastern edition of Microsoft Windows edition supports several double byte character sets.

Why not try to develop a double byte character set system for Tamil also ? On the face of it, it might be a ridiculous suggestion. But in reality it is not. We are only now in the process of standardisation. We should have an open mind and discuss the pros and cons of this system and if the pros outweigh the cons we should definitely adopt the double byte system.

The single byte based Tamil fonts have been in use for over a decade and we have had no problem except that of standardisation. In fact we are all very comfortable with this system and all readily available software packages are able to handle Tamil easily.

Then Why a double byte system? Let me now answer that question. Mind you the thoughts that I am now making have not been researched but only serve to kindle us to think with an open mind before making a final decision on Standardisation.

Till a few years back the use of computers in India was mainly limited to business houses whose requirement of Tamil, if at all any, was limited only to the word processing environment. But the present policy of the Government of India and several State Governments has brought the computer within reach of the common man. In this changed scenario we can now think of extending the use of Tamil to database applications also. As we are aware sorting and searching processes are very vital to these application. The present single byte encoding schemes can be used provided we have a separate sorting algorithm for Tamil.

The computing world is now moving towards UNICODE which has established a global standard for text information interchange. The UNICODE standard only aims at having a single comprehensive coding scheme for all the different world languages and is not concerned about transliteration of these languages as envisaged by the ISCII standard for Indic languages.

The ISCII standard specifies only a 128 character set for each of the Indic languages and all the disadvantages of the 128 character set as discussed earlier applies to the ISCII standard. The ISCII standard gives unique location for each of the vowel, consonant, and modifier characters and leaves the screen rendering process to the software / Hardware. This has been adopted to make transliteration between the Indian languages possible. But transliteration is not perfect because of the inherent differences in the languages. Even presuming without admitting that the available transliteration is sufficient, the use of transliteration is very insignificant and the standardisation process should not consider this as a major criteria. In cases were transliteration is required separate transliteration software can be developed. The simple solution will be to disregard the ISCII standard and evolve a separate standard for Tamil. But the problem does not end here. Unfortunately UNICODE has adopted the ISCII standard and allotted only 128 locations for Tamil (U+0B80 to U+0BFF).

It will not be out of place to mention that several languages have been allotted more than 256 locations in the UNICODE structure. Examples are Ethiopian (U+1200 to U+137F) and Canadian syllabics (U+1400 to U+167F) in addition to of course the CJK languages.

This being the case why can't we get the required number of locations in the UNICODE structure to accommodate each of the 256+Tamil characters individually and not rely on the advantage of creating these characters by adjacent placement of two or more glyphs. By doing this we have a simple one to one mapping between the glyphs and each of the 256+tamil characters. This will greatly simplify, the sorting and searching process in database applications also.

The next question that comes to our mind is what about software? More and more applications are now becoming UNICODE compliant and the day is not far off when almost all software will become UNICODE compliant. Hence the question of Software compatibility is only a temporary phase.

Let us try to utilise the big opportunity given by the UNICODE structure to our advantage. Let us not be myopic and still think that only a maximum of 128 or 256 character sets are available to us and that we have to cut our feet to match our shoes.

**Conclusion :** The above discussion clearly leads to the conclusion that for the present we should standardise on a Single Byte Monolingual font system which facilitates easy sorting and searching. The existing Monolingual font systems should be analysed from this point of view and a suitable one could be adopted as the standard. A possible starting point could be the Chennai code Version 1.4 which has been proposed by Dr.V.Krishnamoorthy. These fonts should also without doubt ensure a high quality of printed text which is of utmost importance not only to the Print media but also to the cause of Tamil in general.

But the future of computing lies in UNICODE and we should try to exploit the wonderful opportunities thrown open by it. We should seriously think about the Double byte encoding scheme as it will greatly help in database applications and we should certainly not try to live within the confines of the 256 character limitation.

## Annexure - A
## Problems associated with Kerning

1. Kered characters loose their aescthetics. As can you clearly see below a single "kokki" glyph will not suffice. The gap between the character and the glyph will be inconsistent and in fact for the letters 'ma' and 'sha' it will almost coincide with the letter itself while for the letters 'ka' and 'sa' it will be far apart.

மி கி ஷி சி

2. During justification the glyphs of the kerned charactrer will be pulled apart. As you can see below the Tamil letter 'sa' and the 'kokki' are pulled apart in the first row. The letter 'ma' looks ok in the first row but in the second row it is pulled apart.

மாமி வாசி மாமி வாசி மாமி
வா மாமி மாமி மாமி மாமி
மாமி

3. Because of kerning, characters such as those shown below in the first row cannot be designed as you have a choice of only one 'kokki'. With kerning you can only have characters as shown in the second row.

பி யி ஷி தி தி
பி யி ஷி தி

# ANNEXURE-B

## ENCODING SCHEME FOR TAMIL BASED ON 512 CHARACTER SET

| DECIMAL CODE | CHARACTER | DECIMAL CODE | CHARACTER | DECIMAL CODE | CHARACTER | DECIMAL CODE | CHARA( |
|---|---|---|---|---|---|---|---|
| 0 - 31 | ASCII | 173 | ஃ | 232 | ழி | 291 | நூ |
| 32 | <Space> | 174 | க | 233 | ளி | 292 | பூ |
| 33 | ! | 175 | ங | 234 | றி | 293 | மூ |
| 34 | " | 176 | ச | 235 | னி | 294 | யூ |
| 35 | # | 177 | ஞ | 236 | ஹி | 295 | ரூ |
| 36 | $ | 178 | ட | 237 | ஸி | 296 | லூ |
| 37 | % | 179 | ண | 238 | ஜி | 297 | வூ |
| 38 | & | 180 | த | 239 | ஷி | 298 | ழூ |
| 39 | ' | 181 | ந | 240 | கீ | 299 | ளூ |
| 40 | ( | 182 | ப | 241 | ஙீ | 300 | றூ |
| 41 | ) | 183 | ம | 242 | சீ | 301 | னூ |
| 42 | * | 184 | ய | 243 | ஞீ | 302 | ஹூ |
| 43 | + | 185 | ர | 244 | டீ | 303 | ஸூ |
| 44 | , | 186 | ல | 245 | ணீ | 304 | ஜூ |
| 45 | - | 187 | வ | 246 | தீ | 305 | ஷூ |
| 46 | . | 188 | ழ | 247 | நீ | 306 | கெ |
| 47 | / | 189 | ள | 248 | பீ | 307 | ஙெ |
| 48 | 0 | 190 | ற | 249 | மீ | 308 | செ |
| 49 | 1 | 191 | ன | 250 | யீ | 309 | ஞெ |
| 50 | 2 | 192 | ஹ | 251 | ரீ | 310 | டெ |
| 51 | 3 | 193 | ஸ | 252 | லீ | 311 | ணெ |
| 52 | 4 | 194 | ஜ | 253 | வீ | 312 | தெ |
| 53 | 5 | 195 | ஷ | 254 | ழீ | 313 | நெ |
| 54 | 6 | 196 | கா | 255 | ளீ | 314 | பெ |
| 55 | 7 | 197 | ஙா | 256 | றீ | 315 | மெ |
| 56 | 8 | 198 | சா | 257 | னீ | 316 | யெ |
| 57 | 9 | 199 | ஞா | 258 | ஹீ | 317 | ரெ |
| 58 | : | 200 | டா | 259 | ஸீ | 318 | லெ |
| 59 | ; | 201 | ணா | 260 | ஜீ | 319 | வெ |
| 60 | < | 202 | தா | 261 | ஷீ | 320 | ழெ |
| 61 | = | 203 | நா | 262 | கு | 321 | ளெ |
| 62 | > | 204 | பா | 263 | ஙு | 322 | றெ |
| 63 | ? | 205 | மா | 264 | சு | 323 | னெ |
| 64 | @ | 206 | யா | 265 | ஞு | 324 | ஹெ |
| 65-76 | Tamil Numerals | 207 | ரா | 266 | டு | 325 | ஸெ |
| 77-90 | Blank | 208 | லா | 267 | ணு | 326 | ஜெ |
| 91 | [ | 209 | வா | 268 | து | 327 | ஷெ |
| 92 | \ | 210 | ழா | 269 | நு | 328 | கே |
| 93 | ] | 211 | ளா | 270 | பு | 329 | ஙே |
| 94 | ^ | 212 | றா | 271 | மு | 330 | சே |
| 95 | _ | 213 | னா | 272 | யு | 331 | ஞே |
| 96-122 | Blank | 214 | ஹா | 273 | ரு | 332 | டே |
| 123 | { | 215 | ஸா | 274 | லு | 333 | ணே |
| 124 | \| | 216 | ஜா | 275 | வு | 334 | தே |
| 125 | } | 217 | ஷா | 276 | ழு | 335 | நே |
| 126 | ~ | 218 | கி | 277 | ளு | 336 | பே |
| 127-160 | Blank | 219 | ஙி | 278 | று | 337 | மே |
| 161 | அ | 220 | சி | 279 | னு | 338 | யே |
| 162 | ஆ | 221 | ஞி | 280 | ஹு | 339 | ரே |
| 163 | இ | 222 | டி | 281 | ஸு | 340 | லே |
| 164 | ஈ | 223 | ணி | 282 | ஜு | 341 | வே |
| 165 | உ | 224 | தி | 283 | ஷு | 342 | ழே |
| 164 | ஊ | 225 | நி | 284 | கூ | 343 | ளே |
| 167 | எ | 226 | பி | 285 | ஙூ | 344 | றே |
| 168 | ஏ | 227 | மி | 286 | சூ | 345 | னே |
| 169 | ஐ | 228 | யி | 287 | ஞூ | 346 | ஹே |
| 170 | ஒ | 229 | ரி | 288 | டூ | 347 | ஸே |
| 171 | ஓ | 230 | லி | 289 | ணூ | 348 | ஜே |
| 172 | ஔ | 231 | வி | 290 | தூ | 349 | ஷே |

| DECIMAL CODE | CHARACTER | DECIMAL CODE | CHARACTER |
|---|---|---|---|
| 350 | கை | 412 | ஸோ |
| 351 | ஙை | 413 | ஜோ |
| 352 | சை | 414 | ஷோ |
| 353 | ஞை | 415 | கௌ |
| 354 | டை | 416 | ஙௌ |
| 355 | ணை | 417 | சௌ |
| 356 | தை | 418 | ஞௌ |
| 357 | நை | 419 | டௌ |
| 358 | பை | 420 | ணௌ |
| 359 | மை | 421 | தௌ |
| 360 | யை | 422 | நௌ |
| 361 | ரை | 423 | பௌ |
| 362 | லை | 424 | மௌ |
| 363 | வை | 425 | யௌ |
| 364 | ழை | 426 | ரௌ |
| 365 | ளை | 427 | லௌ |
| 366 | றை | 428 | வௌ |
| 367 | னை | 429 | ழௌ |
| 368 | ஹை | 430 | ளௌ |
| 369 | ஸை | 431 | றௌ |
| 370 | ஜை | 432 | னௌ |
| 371 | ஷை | 433 | ஹௌ |
| 372 | கொ | 434 | ஸௌ |
| 373 | ஙொ | 435 | ஜௌ |
| 374 | சொ | 436 | ஷௌ |
| 375 | ஞொ | 437 | க் |
| 376 | டொ | 438 | ங் |
| 377 | ணொ | 439 | ச் |
| 378 | தொ | 440 | ஞ் |
| 379 | நொ | 441 | ட் |
| 380 | பொ | 442 | ண் |
| 381 | மொ | 443 | த் |
| 382 | யொ | 444 | ந் |
| 383 | ரொ | 445 | ப் |
| 384 | லொ | 446 | ம் |
| 385 | வொ | 447 | ய் |
| 386 | ழொ | 448 | ர் |
| 387 | ளொ | 449 | ல் |
| 388 | றொ | 450 | வ் |
| 389 | னொ | 451 | ழ் |
| 390 | ஹொ | 452 | ள் |
| 391 | ஸொ | 453 | ற் |
| 392 | ஜொ | 454 | ன் |
| 393 | ஷொ | 455 | ஹ் |
| 394 | கோ | 456 | ஸ் |
| 395 | ஙோ | 457 | ஜ் |
| 396 | சோ | 458 | ஷ் |
| 397 | டோ | 459 | லை |
| 398 | ணோ | 460 | ளை |
| 399 | தோ | 461 | னை |
| 400 | நோ | 462 | ணை |
| 401 | போ | 463 | றா |
| 402 | மோ | 464 | னா |
| 403 | யோ | 465 | ணா |
| 404 | ரோ | 466 | ஸ்ரீ |
| 405 | லோ | 467 | க்ஷ |
| 406 | வோ | 468 | க்ஷ் |
| 407 | ழோ | | |
| 408 | ளோ | | |
| 409 | றோ | | |
| 410 | னோ | | |
| 411 | ஹோ | | |

# PROPOSAL FOR STANDARD ENCODING OF TAMIL CHARACTERS

*Sinnathurai Srivas*

**Abstract**

*This is a proposal for a Standard Encoding of Tamil Characters for use in Electronics, Computer and Communication industry. Inorder to formalise the standardisation of Tamil characters, extensive work have been done by various enthusiasts, Voluntary Organisations and Government Departments. A significant amount of work have been done by 1/ volunteers at tamil.net mailing forum , 2/ TamilNetXX, 3/ The Tamilnadu Computer Standardisation Committee, and various other individuals, organisations, business establishments, universities, etc.*

*This proposal is based on the factors considered at www.tamil.net mailing list forum.*

*The followings as a proposal is to be submitted to The Tamilnadu Computer Standardisation Committee and to TamilNet'99 for their evaluation as a Standard Encoding for Tamil.*

## The Proposal

1/ LOGICAL Sorting in Tamil

2/ Kerning and Aligning in Tamil

3/ The Present and The Future of Tamil Characters

4/ Keyboard Drivers

5/ The Tamil Code Page in ASCII map (ASCII-TC)

6/ The Unicode and Tamil (UicodeTC)

## 1/ LOGICAL Sorting in Tamil

For a reader/writer of Tamil language, even though it looks very simple to recognise a Mey Ezuththu, a very close look at the script would reveal that it does have a complex set of ways to make up the Uyr-Mey Ezuththukkal.

They are:

Independent representation of uhara and uuhaara Uyr-Mey Ezuththukkal.

கு கூ சு சூ.......

Frontal modification by combus to obtain ehara and eehaara Uyr-Mey Ezuththukkal.

தெ, தே

Frontal modification by combus and rear modification by kaNNa aytham to obtain eehara and ohaara Uyr-Mey Ezuththukkal.

தொ, மோ, .......

Modifier oTai combu providing ohara modification, while modifier iTedai combu providing oohaara modification. Similar operations for ehara and eehaara Mey Ezuththukkal .

தொ (த்ஒ), தோ (த்ஓ)

தெ (த்எ), தே (த்ஏ)

Back sided modifications to obtain ihara and iihaara Mey Ezuththukkal.

பி,நீ,......

Ahara-Mey Ezuththukkal represented by core Mey Ezuththukkal and invisible (manathil vykkum) aharam.

க், ....... (க்அ) (க+)

Mey Ezuththukkal represented by Core Mey Ezuththukkal and the Kuttu modifier.

ழ், .......

Prime function L also acting as secondary function L2 and prime function oTaicombu also having secondary function in ouhara Mey Ezuththukkal.

கௌ, ளௌ, .....

These make the job of designing an encoding more challenging.In the following sections, this proposal suggests the elements that can go into the Tamil Encoding Table based on the following criteria:

- Direct, wrapper-less Logical Sorting enabled encoding
- Reduce cost and to make programming simple
- Reduced Set of Glyphs
- To keep the important glyphs in secured slots of the table and to enable Logical Sorting.

The sorting is to be done using standard ASCII (**ISO 8859-X**)/(ANSI!) compliant coding to obtain a Logical sorting capability for Tamil.

The structure of Logical sorting as defined below has been elected after much thought devoted to technical and usability factors. As it is based on the logic of Tamil writing this may be used as the de-facto protocol for Tamil sorting or it could co-exist with the Grammatical sorting for Tamil. The logic has it that it is more user friendly than the Grammatical method. The advantageous position of the logical sorting is that it co-exists with the internationally agreed sort order of ASCII code, where by facilitating the sorting algorithms bundled with major software to be invoked for sorting in Tamil, drastically minimising the need for add-on software. The proposed logical sort order is to be as in Table 2.8.1. ( To view the sort order of character codes see "t99XL1.htm" Table 1.2. and Table 1.3.)

**Table 2.8.1**
**Proposed logical sort order for Tamil characters**

| | |
|---|---|
| | அஆஇஈஉஊஎஏஐஒஓஔஃ |
| | ெ ே ை ஙுஙூஞுஞூ ள2 க்ர்ணo ஃ ா ா ு |
| | கக்குகூ ஙங்ஙுஙூ சச்சுசூ ஞஞ்ஞுஞூ டட்டிடீடுடூ<br>ணண்ணுணூ தத்துதூ நந்நுநூ பப்புபூ மம்முமூ<br>யய்யுயூ ரர்ருரூ லல்லுலூ வவ்வுவூ ழழ்ழுழூ<br>ளள்ளுளூ றற்றுறூ னன்னுனூ ணண்ணுணூ |

Taking the sorting as a criteria the above is achieved by allocating the upper ASCII slots with their sort (internationally agreed practice) values equalising to the sort order as given above.

In evaluating the other factors that influences the selection of ASCII slots, what effect this will have on the overall function. The important factor that will be disturbed by the influence of sorted code order is that, it will be necessary to assign more of the ASCII slots identified as insecure for character usage.

Now come the question, lets look at all the criteria that influence the end system design. One of the major criteria is logical sorting. What are the other major or minor criteria that can be weighed and analysed to get this functionality incorporated. On careful analysis it becomes necessary to compare and contrast the other major factor that influences overall system, the minimised kerning complexities. The minimised kerning requirement, particularly of the no kerning - full representation criteria used for all Meys become a necessity to be re-evaluated.

## 2. Kerning and Aligning

### 2.1. Kerning and Aligning of Ahara-Meys with Kuttus.

Let us re-visit the factors considered before allocating codes for all ahara-meys independently and all meys independently.

i) Except for vertical kerning, kerning/aligning of ihra and iiharam meys are not complex than kerning/aligning of kuttus. A perfect kerning of any denominations are only available for professional print houses where even the normal character combinations go through intense kerning process as part of typography.

ii) DTP is another profession where the printing is at the low end of technological adequacies. Judging from the reports coming from the DTP user base some equipment, fonts and necessary application packages are capable of handling the existing available technologies at this market level to perfection while others report having undesirable results either at equipment level, or at font level, or at application packages level. Notably DOS based users are of the latest variety.

iii) Display systems are another of the important component part of this investigation. Again here the major reluctance was with the ability to handle raw text handlers alas text editors. From the emerging support for Tamil in text handling, it is again some equipment or fonts or the text handlers are producing perfect results while some reports not satisfactory in handling either aligning or kerning. Here again DOS based systems are sighted as problematic.

iv) Other forms of display devices such as POS and LCD based displays need further investigation, while bearing in mind that the Unicode system will be in use while this type of systems become widely used in Tamil arena.

Advantages

(1) Sorting is achieved without the aid of add-on software ( minimised cost and complexion).

(2) Character in the ASCII domain slots are allocated at even more secure codes

(3) Freeing of additional codes enables additional functionality to be invoked.

Disadvantages

(1) Kerning is invoked, particularly of uyr-meys.

(2) Quality of DTP and screen display may suffer, due to kerning deficiencies (increased cost, inadequate technologies)

Having considered this in accordance with "The encoding standard must be such as to be readily implemented in most of the widely used computer platforms (UNIX, Windows and Mac)" and other factors, it is desirable to invoke aligning and kerning for kuttus to output meys for ASCII based coding.

## 2.2. The Kerning and Aligning of some uuhaarameys

நூ, தூ, ஞூ, ணூ, ஙூ, பூ, யூ, வூ, ஙூ, (ா ூ)

As the aligning and kerning of the uuharameys are relatively simple task compared with iharameys, iihaarameys and meys these characters will also not be represented as integrated ones but be modified with their pairing kaNNa modifiers.

## 3. The Present and The Future of Tamil Characters

The design also takes into consideration the future of Tamil writing by reserving code positions for adding support for researchers. The design has a future proof capability, providing support for anticipated changes, if they are to be invoked in a controlled manner.

### 3.1. Additional and Future Support.

The design of code allocation provides support for the researchers to take their slow and steady implementation for the good of future of our generations.

As part of these action it is not the allocating of code positions at random that will facilitate the forward march, this allocation of slots also have been studied and implemented in accordance with the inbuilt sorting capability.

| | |
|---|---|
| Reserved for Future expansion | [illegible] |
| Provide additional Support | x° x௸ ஹ(க்ஷ) கூ ஸஷக்ஷ டபு (யு) ங |
| Support tools for Research | ◌ு◌ு ◌ு◌ு ( நிபுந்தனை) |
| Support tools for Research | ◌் ◌்◌் ◌் (அ்ம்மா அ்ன்னை) |
| Support tools for Research | ◌்◌் (க்ஷ்=க்ஷ), ப்றி) |
| Provide Optional support | ஸ்ரீஜெ☪✝ ... x x x x |
| Provide Optional Currency Symbol | ரூ ச |

## 4. The Tamil Code Page in ASCII map: ASCII – TC

The Tamil code page on ASCII character map is given in Table 2.8.2.

## Table 2.8.2 ASCII-TC Proposed character codes for Tamil.

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 |
| | ! | " | # | $ | % | & | ' | ( | ) | * | + | , | - | . | / |
| 48 | 49 | 50 | 51 | 52 | 53 | 54 | 55 | 56 | 57 | 58 | 59 | 60 | 61 | 62 | 63 |
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | : | ; | < | = | > | ? |
| 64 | 65 | 66 | 67 | 68 | 69 | 70 | 71 | 72 | 73 | 74 | 75 | 76 | 77 | 78 | 79 |
| @ | A | B | C | D | E | F | G | H | I | J | K | L | M | N | O |
| 80 | 81 | 82 | 83 | 84 | 85 | 86 | 87 | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 |
| P | Q | R | S | T | U | V | W | X | Y | Z | [ | \ | ] | ^ | _ |
| 96 | 97 | 98 | 99 | 100 | 101 | 102 | 103 | 104 | 105 | 106 | 107 | 108 | 109 | 110 | 111 |
| ` | a | b | c | d | e | f | g | h | i | j | k | l | m | n | o |
| 112 | 113 | 114 | 115 | 116 | 117 | 118 | 119 | 120 | 121 | 122 | 123 | 124 | 125 | 126 | 127 |
| p | q | r | s | t | u | v | w | x | y | z | { | \| | } | ~ | □ |
| 128 | 129 | 130 | 131 | 132 | 133 | 134 | 135 | 136 | 137 | 138 | 139 | 140 | 141 | 142 | 143 |
| □ | ı | கு | கூ | ı | ... | 1˙ | 2ர் | ¨ | ா | ரு | ஹ | ர | ı | □ | ı |
| 144 | 145 | 146 | 147 | 148 | 149 | 150 | 151 | 152 | 153 | 154 | 155 | 156 | 157 | 158 | 159 |
| ı | ‘ | ’ | “ | ” | 3ண் | ௐ | ☪ | ͵ | TM | ல | ஸ | ரு | ı | □ | ண |
| 160 | 161 | 162 | 163 | 164 | 165 | 166 | 167 | 168 | 169 | 170 | 171 | 172 | 173 | 174 | 175 |
| | ı | ச | ந | ~ | ~ | ! | ந | ÷ | © | ˧ | ெ | ட | க்ஷ | ® | ° |
| 176 | 177 | 178 | 179 | 180 | 181 | 182 | 183 | 184 | 185 | 186 | 187 | 188 | 189 | 190 | 191 |
| ı | ஸ்ரீ | ஞு | ள | . | ை | ஜ | 4ŕ | ˩ | ஙு | த | ே | ா | ் | ı | ^ |
| 192 | 193 | 194 | 195 | 196 | 197 | 198 | 199 | 200 | 201 | 202 | 203 | 204 | 205 | 206 | 207 |
| ஆ | அ | இ | உ | ஈ | ஊ | எ | ஏ | ஒ | ஓ | ஔ | க | சு | ஙு | ஞ | டி |
| 208 | 209 | 210 | 211 | 212 | 213 | 214 | 215 | 216 | 217 | 218 | 219 | 220 | 221 | 222 | 223 |
| ஐ | டு | நு | து | பு | மு | ᒧ | ஷ | ய | ழு | ழ | ளு | ற | ன | வ | லு |
| 224 | 225 | 226 | 227 | 228 | 229 | 230 | 231 | 232 | 323 | 324 | 325 | 326 | 327 | 328 | 329 |
| ரி | ிா | ீ | ிா | ி | ு | ுா | ஹ | ஸ் | ளா | ஃ | கு | கூ | ச | ட | டீ |
| 240 | 241 | 242 | 243 | 244 | 245 | 246 | 247 | 248 | 249 | 250 | 251 | 252 | 253 | 254 | 255 |
| ர | டே | ப | ந | ய | மூ | ம | கு÷ | யு | ள | மு | ளு | று | னு | வு | ணு |

**Keyboard help for Testing the Tamil_Ava99_v1 font.**

This keyboard was created from the Keyman keyboard Programm, Version2 c Written by Sinnathurai Srivas 2 Jan 1999.

This program is written only to help with the initial typing for the font Tamil_Ava99_v1. This key strokes or k/b layout does not confirm to any standard and not intended as such. Do not distribute this k/b program for general use.

NAME "*Tamil99_102*", HotKey Ctrl+Alt+Z, Keyman Version 3.2

1. ; > kuttu
2. keep pressing ; to get various kuttus
3. ` > to get mey-aytham
4. keep pressing ` to get various aytham
5. ^ > to get uyr-aytham
6. keep pressing ^ to get various aytham
7. A, AA, I, II, U,UU,E,EE,B,X to get அ,ஆ,இ,ஈ,உ,ஊ,எ,ஏ,ஐ,ஔ

8. hா, Hஃ, bை, oே, iி, iiீ, eெ, [ா, ]ு

9. u for uhara meys, uu for all other uuhara meys

10. qwerty p ஞநறெதய ப, WwT ஙநர

11. sdfghjkl சடfgாஞகல, SDFGHJKL SḌFGஃJKள

12. zxcvbnm ழளசவனைம, ZXCVBNM ZஔCVஜணM

In Tavultesoft Keyboard Manager program Version3.2, Load Tamil99_102 to run the program. Actual file name of the k/b driver is AvaKe4.kmn

# iDNS, a DNS SYSTEM WITH MULTILINGUAL SUPPORT

*Mr Ching-Hong Seng & Dr Tin-Wee Tan*
*Singapore*

**Abstract**

*The Internet is based on IP technology. Packets are routed using IP address which is in numerical form, for example 137.132.19.110. While this is useful for routing purposes, it is difficult to remember. Thus, we have a domain name system or DNS which translate domain names, eg www.nus.edu.sg to IP addresses and vice versa.*

Navertheless, the current DNS only allows a very limited set of characters. While this is not a problem for most English speaking countries, this is not so for many countries in the Asia Pacific region where its own national language is more used.

Thus, with the internationalisation of the Internet, and the proliferation of non-English content, there is a growing need for a DNS system which is able to handle multilingual and multiscript features.

In addition, with the development of Unicode which encompass most languages in the world, including Tamil, it is high time a DNS system is designed for non-English speaking countries.

The iDNS aims to address this issue by

a. Providing a mechanism for the DNS system to adopt multilingual character set thus allowing non-English language domain names.

b. Retains compatibility with the current DNS standards.

c. Transparently transform different language encoding to a universally adopted format Unicode.

d. Minimum changes on the client/server
   — easy installation for the system administrator
   — allowing end-users to immediately use multilingual domain names

## References

RFC1034, P. Mockapetris, Nov. 1987
Domain Names - Concepts and Facilities

RFC1035, P. Mockapetris, Nov. 1987
Doman Names - Implementation and Specification

draft-duerst-dns-i18n-02.txt, M. Duerst, July 1988
Interationalization of Domain Names

draft-skwan-utf8-dns-01.txt, Stuart Kwan/James Gilroy, Mar 1998
Using the UTF-8 Character Set in the Domain Names System

# COMPUTING WITH TAMIL
## Use of the vernacular in designing user interfaces.

*Dr. Kalyanakrishnan, Systems Development Laboratory, Department of Computer Science, Indian Institute of Technology, Chennai 600 028.*

## Foreword:

Effective dissemination of knowledge and information is the key to the well being of any society. The assimilation of knowledge, be it for leading a healthy life or for furthering the cause of education, starts at the school and continues through other means which includes self learning as well. The computer has come to be accepted as an effective tool for collecting, processing and disseminating information in all walks of human life. There is an emerging need to utilize this tool on a large scale to bring the benefits of Information Technology to the people as it can play a vital role in basic education and public administration. Knowledge of English on the part of the person learning to use a computer, has been a requirement all along and this has come in the way of effective implementation of IT solutions in the State as well as in the country. Over the years, many successful attempts have been made to provide computers with user interfaces in Tamil and other Indian Languages. Most of the solutions provided have established the feasibility of interaction with a computer in local languages and have led to the availability of quality word processing and publishing software. Yet, a number of different applications that should form an essential part of the computing environment in Tamil have not been viewed with seriousness on account of the technical challenges faced in dealing with local languages on computers.

This paper presents in the form of a short tutorial, the important technical issues which require consideration before any IT solutions or applications supporting a user interface in Tamil can be implemented. Some recommendations have also been effected in respect of the standardization envisaged for the use of computers in Tamil.

### *Computing with Tamil: An overview.*

The essential idea here is that people will be able to use computers to deal with information given in Tamil. This means that input and display are both supported in Tamil and will not require the use of English. Thus the problem, if approached in a manner which does not require the knowledge of English on the part of the user. assumes significance in our country.

Any system of computing has to meet the basic requirement that it be able to accept, process and display text based information. Ultimately, it is text that relates directly to human communications and hence it is quite important that a computer supporting Tamil based interfaces should in the least, be able to handle data entry and display in Tamil, if not text processing. Electronic information processing is not critical for displaying and printing text since very little processing is involved on the entered data except formatting. The techniques available for doing this with English may be easily adapted to handle Tamil. However, true computing involves the collection of varied items of information, pooling them together and processing the same to produce results consistent with user requirements, such as searching for information in a data base. Applications involving records in public institutions are examples of text processing in a large scale.

A minimal computing environment should therefore support the concept of internal representation of information in Tamil in a meaningful way as to permit a computer program to interpret the information and process the same. Thus there are three essential components to a computing system in Tamil.

1. Data entry of text in Tamil.
2. Display or printing of text in Tamil.
3. Processing of the text to suit user requirements.

Approaches to dealing with all the three have been standardized for the western languages. This standardization has been effected through the concept of character codes and fonts. This standardization has allowed English to be handled in a uniform manner on all computers which use the standard ASCII code for all text processing.

It turns out that data entry, display or printing of Information in Tamil can be effected using the standardization accepted for English, by designing suitable fonts for the display of Tamil letters and using conventional word processing software to effect data entry. This way, word processing and Desk Top Publishing programs are fooled into believing that they are handling Roman text (English). Thus, with a suitable font and the use of the Macro facilities of word processors which allow a sequence of key strokes to be interpreted together, data entry and printing in Tamil may be comfortably handled. Many commercial packages take this approach to providing DTP facilities in Tamil. Unfortunately, such solutions cannot provide meaningful support to the real text processing in Tamil which has to be handled purely from a linguistic or language related point of view.

To perform any meaningful text processing in Tamil (or for that matter any Indian language) one has to work with the syllabic structure of the language and not merely with the vowels and consonants. This is a fairly complex issue as it involves individual identification of hundreds of characters which result through the combinations of the basic vowels and consonants. The phonetic aspect of Tamil makes it necessary for one to relate the written form to the linguistic units (refered to in Tamil as உயிர்மெய் எழுத்துக்கள் ) and there lies the basic difficulty in deriving a suitable internal representation for the text.

A number of available approaches to dealing with Tamil on the computer have not looked at the text processing requirement with much seriousness, and have concentrated primarily on data entry and printing only. As a consequence, one sees a multitude of fonts for Tamil, each with its own accompanying data entry software. One need not be alarmed at this as most of these offer working solutions capable of producing quality printouts or web displays on the internet. The need to look at standardization arises only when one considers the issue of text processing. While such processing is in principle provided for with the internal representations used with fonts, it is nevertheless a complicated task as it bears no relationship to the linguistic aspects of Tamil.

The attempts at the Center for the Development of Advanced Computing have resulted in some standardization in respect of coding and the internal representations of text in Indian languages. This approach, which became popular through the GIST technology, has fairly severe limitations in practical use, especially for Tamil. The CDAC approach does however allow, a number of DOS based applications to work with Indian scripts. There are many difficulties in using the system because of the restriction on the platform and the hardware approach in using the GIST card. Yet, one must accept that the ISCII coding scheme used by the software does allow some computing in Indian languages.

To summarize, computing with Tamil should address the problem of text processing. In the western world, the writing systems are based on a small set of symbols and letter shapes. The software caters to the standardized representations of the text through the ASCII code. The letters of Tamil do not lend themselves to such interpretations, for they represent sounds at the level of a syllable derived from a basic set of vowels and consonants. Each syllable is unique from a linguistic point of view and any effort at standardization or electronic representation of text should necessarily conform to linguistic requirements.

## *Some statistics on the characters used in Tamil.*

There are twelve basic vowels in Tamil (உயிரெழுத்துக்கள் ) and eighteen consonants (மெய்யெழுத்துக்கள் ) divided into three groups. The ஆய்த எழுத்து ஃ is viewed as the thirteenth vowel and each consonant in its intrinsic form (without any vowel) is viewed as the thirteenth representation of the consonant, where each representation is a combination of the consonant and a vowel. Thus there are 13+18+(18*13) different combinations which are significant linguistically.

Added to these are six consonants from Sanskrit which are used fairly regularly these days, though one does not see them in ancient texts. The thirteen combinations of each of these with the basic vowels is also to be represented. Tamil has its own representation for the numerals.

While writing, a consonant vowel combination is not always subject to the same rule of adding a vowel extension, for combinations with the vowels உ and its long form, alter the shape of the consonant itself.

Though punctuation marks are rarely seen in older texts, there is a need to use some of the special symbols and punctuation marks if we have to use the language for preparing educational material and scientific text.

Though Tamil comes under the general category of phonetic languages, the pronunciation of some intermediate letters in words is context dependent and from a linguistic angle quite important. Hence some mechanism to distinguish the different sounds produced by the same letter will be helpful in the preparation of educational material for teaching Tamil.

Lastly, the writing system for Tamil has evolved over a number of centuries and it is necessary to display ancient tamil text in the same scripts used in old manuscripts and stone inscriptions.

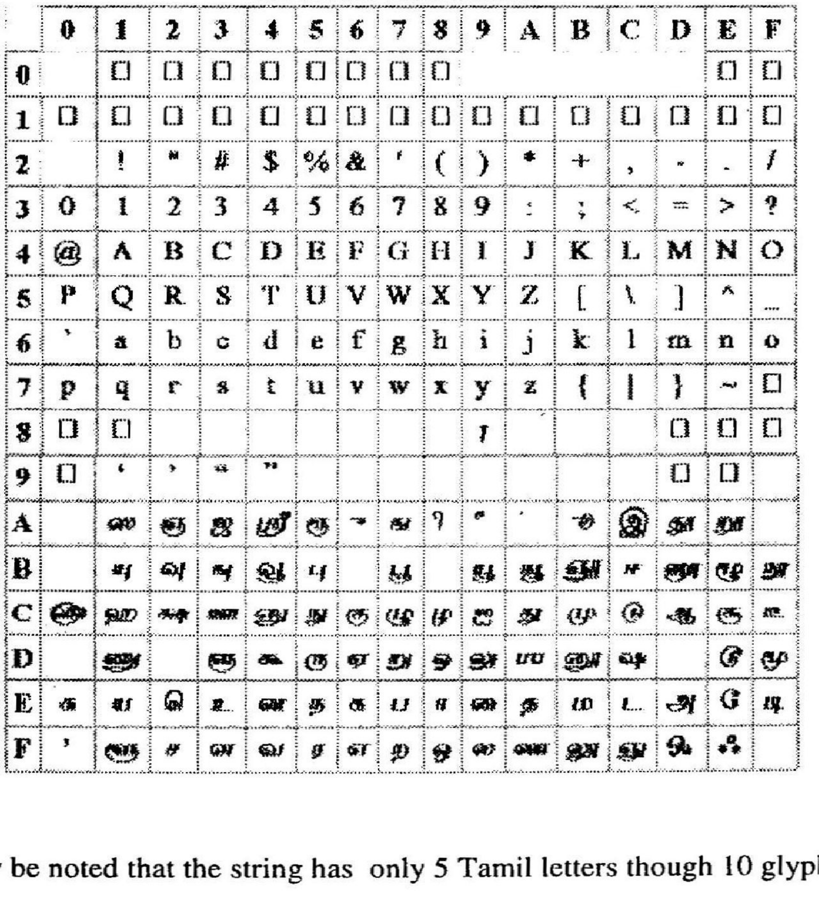
## *A brief overview of the method of using Tamil with Fonts.*

Many of the solutions provided in the past for dealing with Tamil on computers have relied on the use of special fonts. A font is essentially a set of basic shapes which may be shown individually or in combinations to display the letters of a script. If we look at a string such as தமிழ்மொழி , the individual shapes which build up the string are,

த ம ி ழ ் ெ ம ா ழ ி

There are 10 shapes here (each is known as a glyph) but only seven of them are distinct. The glyphs within a font are arranged in some order and inside the computer, the shapes to be displayed are given in terms of the positions occupied by the glyphs in the fonts. Typically a font will support up to about 200 glyphs. Shown below is the font used in displaying the Tamil string above. The numbers indicated following this line correspond to the codes of the glyphs used in the string. (234, 235, 168, 200, 208, 226, 235, 232, 200, 168)

INDOWEB-Tamil-Kambar



It may be noted that the string has only 5 Tamil letters though 10 glyphs are used, 7 being distinct.

The method most often recommended for representing information in Tamil is the use of glyph codes for the letters. This works alright for a given font and to some extent string processing may also be attempted. However the dependence of the codes on the font is a major deterrent to using this approach. The glyphs required to display a string are generated when the keystrokes are effected. For this, the keys on the ascii keyboard are mapped into the tamil letter whose glyph is specified through the ascii code for the letter. Some combinations will require two keystrokes but this is alright since the second keystroke will correspond to a vowel extension.

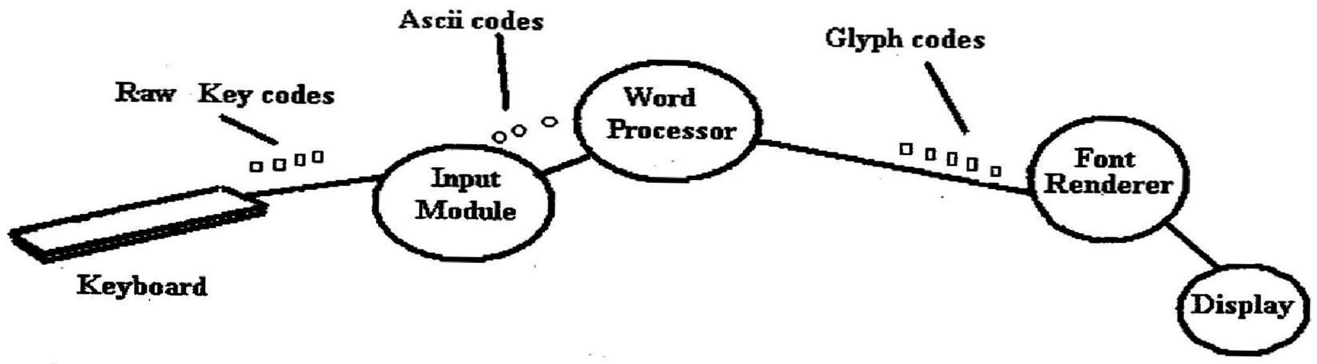
Inspite of its simplicity, this method is quite painful in practice, for it is not always easy to remember the mapping between the Tamil letter and the key. Worse still, some combinations require three keystrokes to be executed whereas other combinations may be handled through just one keystroke. An example of this is the difference between keyboard entries for தி and து. Depending on the font design, the number of glyphs and therefore the number of keystrokes will vary. This is not a useful approach for general acceptance.

It is thus apparent that Tamil cannot be typed in just as English is. Somehow the consonant-vowel combinations will have to be handled in a font independent manner which also allows a keyboard entry procedure based on the same number of keystrokes for any combination. Some word processors provide special support by tracking the keystrokes and combining them appropriately. Packages from CDAC or word processors such as SRILIPI use this method. Here too, the sequence is decided by the glyph mappings though in the CDAC software, a special mapping known as the INSCRIPT is used but this is not intuitive for those familiar with English. With most software, one has very little choice for the fonts since the data entry method is in some way tied to the use of fonts. No two font designers agree on the glyphs nor their placement within the font. Worse still, fonts are not always compatible across computer systems.

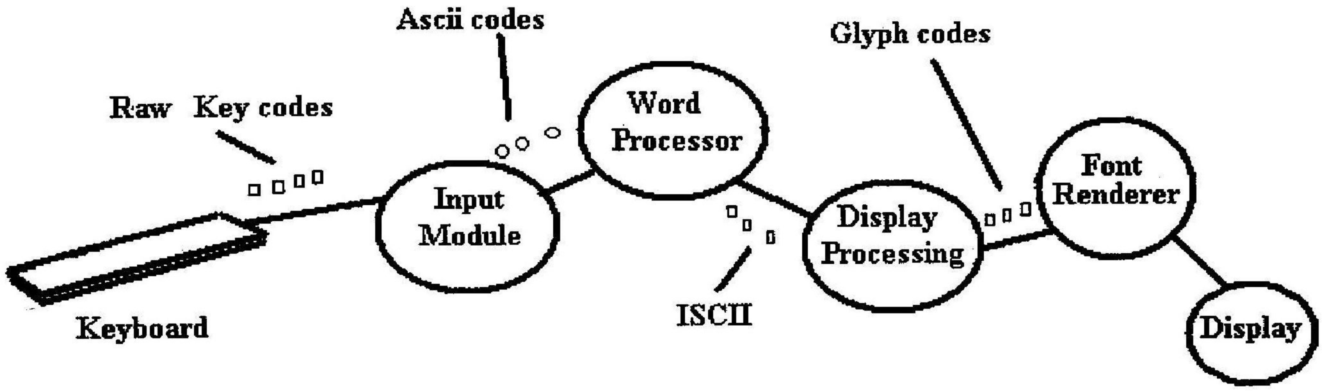
Looking at the problem of keyboard mapping itself, the key to be pressed for a specific letter is fixed by the glyph code. This key may have no phonetic equivalence with the Tamil letter. In fact for many fonts where the designer had intended bilingual use (Roman and Tamil), the glyphs for Tamil are located in the 128-255 range making data entry even more difficult, unless the MACRO features are invoked.

One solution recommended by some designers has been to place the glyphs of the Tamil letters at positions corresponding to the ASCII code of the phonetically equivalent letter in Roman. This makes data entry a bit more intuitive but here too variations occur when dealing with consonant vowel combinations which change the basic shape of the consonant.

The process is illustrated below.



Computer programs from CDAC use a slightly modified approach.




Here, internal storage is not in terms of glyph codes but correspond to the ISCII scheme. The special word processor transforms the keystrokes into appropriate letters and an output module converts the ISCII based internal representation into glyph codes. This method has the advantage that it applies to all the Indian languages. But the scheme itself suffers from some language specific representations.

The ISCII code is an eight bit code that codes only the basic consonants and vowels of the language. Consequently it requires more than one byte to represent a combination though a consonant or a vowel by itself requires only one byte. Thus ISCII amounts to a multibyte variable length code, making the font rendering mechanism quite complex. Nevertheless ISCII is a code that represents basic sounds and hence is quite useful in practice. In implementation however, ISCII has run into some problems for South Indian scripts, especially Tamil.

### *Basic Requirements for coding schemes.*

1. Codes must necessarily correspond to syllables which form the linguistic base for the language. Also the coding must use fixed length codes, even if multibyte. This is the best way to handle complex string processing issues consistent with the phonetic nature of our languages.

2. It will be helpful if glyph positioning within a font for Tamil have some relationship with the internally assigned codes. Such a provision will help in string processing.

3. Codes assigned must conform to the dictionary sorting order for the letters of the language.

4. Codes must also be assigned for the special symbols used in the language, numerals included.

5. Codes should identify the consonant and vowel forming the syllable. This will help in linguistic analysis as well as conversion to other formats. This will also help in speech synthesis.

6. Coding should take into account syllables found in other languages of India so that transliteration into Tamil is easily effected. This will help teach other languages through Tamil.

7. The assignment of codes may also take into consideration the numeric values traditionally assigned to the letters of Indian languages.

8. Codes should have no relationship to the glyphs used to display the letter. This is essential to make sure that the internal representation is independent of the font rendering process. Only then we have the possibility of using the software on different platforms. This recommendation does not really run contrary to the observation made earlier that glyph positioning may be influenced by the internal representation.

## *Recommendations for the design of Fonts.*

1. It will be useful to fix the glyphs in the displayable range for ASCII since data entry can be effected from word processors. This s feasible for Tamil. This way editing can be done on the text string using conventional word processors. Even if placed in the 128-255 range, the glyphs may be located at positions corresponding to the equivalent ASCII in the lower range.

2. It is a good idea to offer a minimum set of glyphs for each script and fix their locations as well. This will make it easier to view the same text in different fonts so long as the text does not incorporate special symbols.

3. Font rendering methods vary across systems and it is a good idea to build fonts with standard encodings so that they may be rendered on different platforms using the same glyph codes. This specification is important for web based applications. The default encoding prefered for Tamil is the latin-8859-1 encoding which is well supported on all the three platforms. Also, the minimal set of glyphs may be used on all the fonts with some advantage.

4. It is a good idea to include punctuation and special symbols in the glyphs for the language. This way, it will be possible to treat special symbols as part of the language itself. Punctuation should include the period, the comma, exclamation mark, the hyphen and the question mark.

5. It will be helpful to provide some special symbols used in ancient manuscripts of Tamil (e.g., shapes for some double consonants).

6. It is inevitable that variable width characters be designed. In the glyphs, it is recommended that the width is made a multiple of some basic unit (say 2pixels). This will help retain vertical alignment in the text without having to resort to special formatting. One glyph may be retained as a special space whose width is an odd number of units. This glyph will be useful in retaining the alignment.

## *Recommendations on the use of the keyboard.*

It turns out that keymappings may be assigned arbitrarily and the processing software can do the required mapping to the internal codes. Though in earlier sections we hinted at mappings that bore some relationship to the font glyphs, it is useful to look at keyboard mappings from the user point of view rather than programming convenience. The following are some of the recommendations.

1. Data entry should be natural and must relate to the letters of Tamil. It should be easy enough to train persons in the use of the standard QWERTY keyboard which is what one will see on all computers.

2. As far as possible, use only the common keys to map the letters. Not all the keyboards will have all the keys seen in the PC keyboards.

3. Standardize the input method for entering a consonant vowel combination.

4. The manual typewriter keyboard (for Tamil) is a choice that should not be ignored. It is an existing standard and many have been trained in it. This keyboard is entirely adequate for modern Tamil writing. The processing software can always use a keyboard filter to transform the sequence of keystrokes into appropriate letters. This may be accomplished through the use of Macros in most word processors.

5. Do not fix or relate keymapping to any fonts though this is an approach that may allow virtually any word processor to handle data entry in Tamil. What is desirable is that the internal representation be exportable to some word processors so that the powerful formatting facilities seen in them may be utilized.

6. Keyboard mappings should not also relate to the codes assigned for the characters. This is to ensure that the internal representation is independent of the system in which we are processing the text. It is always easier to work with standard codes that do not directly relate to any hardware specific aspect of a computer. Tables can be used to relate the internal representation to the glyphs or keyboard mappings thus allowing great flexibility in dealing with the input and the display.

7. Keyboard mappings arrived at on the basis of some studies on the observed frequencies of the letters in normal Tamil writing are no doubt helpful. It turns out that writing styles vary so much that the frequencies seen in ancient texts are quite different from the frequencies in modern texts. It is therefore preferable to look at the thinking process as one types and assign the keymappings based on user recommendations.

## *A brief on the applications to be supported.*

Computing in Tamil, or with Tamil, has to necessarily provide support for some basic applications that will help bring information technology closer to the people. While one might wish to have virtually every popular application run in Tamil, we need to view some applications as being important to begin with. Given below is a list of applications that should be made available in the vernacular at the earliest. Whatever standardization is contemplated in respect of representing Tamil inside the computer, the recommendations should keep these applications in mind and offer viable technical solutions.

1. Data preparation applications.

   * Data entry and printing applications which may include DTP, both for general use and commercial publishing. Large scale newspaper and magazine publishing as well as archiving.

   * Display and dissemination of information through the web. The applications include tools for preparing HTML documents, Search engines, Archiving software, Applications for preserving ancient manuscripts in their original scripts. Generation of on-line references such as dictionaries which may be queried from client systems or web browsers on the net.

2. Text processing applications.

   * Applications involving sorting, indexing and searching of large volumes of data.

   * Data base systems supporting interaction in Tamil. Report generation should be supported in Tamil as well. This will help maintain records in many Government, public and private institutions. The system should be supported on standard Data base packages such as Oracle, Access, Informix etc..

3. Educational applications.

   * Applications involving the teaching of Tamil to the people of other states through their mother tongue. Likewise, learning other languages through Tamil.

   * Preparation of computer based lessons in Tamil to enhance classroom instruction. Lessons should be

interactive as well as web based.

4. Linguistic applications.

   * Analysis and study of the structure of languages.

   * Study of ancient texts from a linguistic angle.

   * Concordance generation.

   * Applications catering to the analysis of sentences, Morphology, word frequency computations etc..

5. Email communications.

   * Software to handle email in Tamil so that the benefits of this wonderful facility are fully utilized.

6. Applications specific to Government and Public Institutions.

   * Maintaining records of text data bases, Police records, Historical data, Minutes of meetings etc., for quick and effective access.

7. Other consumer applications.

   * Accounting and small data base packages supporting user interfaces in tamil.

   * Client server applications supporting user interfaces in Tamil, to work with standard data bases.

8. Applications for the disabled.

   * Computers are proving to be extremely valuable for training disabled children. It is necessary to look into possibilities of using them in the vernacular for the benefit of the disabled.

## Summary

Information Technology, in respect of usage of Tamil on computers on a State-wide scale, should necessarily aim at reaching all the people so as to give them the benefits of this new technological wonder. This is not a simple matter that can be handled by designing fonts or standardizing keyboard mappings. Any approach to dealing with user interfaces in Tamil cannot be based on software solutions that merely cater to data entry and printing, no matter how good the results are, to look at. These solutions will not provide a simple, easy and uniform way of communication that can be used by all people. It is necessary to look at the problem from the angle of information processing for the long term growth of the Tamil language, both for electronic processing and the socio-economic progress of the state. In a wider context, Tamilnadu should be able to share its expertise to provide IT solutions for the rest of the country in their respective languages. Software and business opportunities throughout the country will then come within the reach of the Tamilnadu software industry. We should not loose the opportunity in proposing a viable approach to Information technology in Indian languages.

*எண்ணிய எண்ணியாங் கெய்துப எண்ணியார்*
*திண்ணிய ராகப் பெறின்*

# TOWARDS DEFINING THE DESIGN GOALS WHILE PLACING TAMIL IN THE UNICODE

*Dr. P. Chandra Bose*
*Reader in English*
*Presidency College (Autonomous)*
*Chennai 600 005.*

**Abstract**

*This paper tries to seek and define certain objectives based on the experience of working with various Tamil Desk Top Publishing packages and Word processors while thinking seriously about the present attempts of various forums in finding slots in the UNICODE for Tamil. The approach is basically as that of a user and at the end a researcher in Tamil computing. The objectives are termed Design Goals and an attempt is made to make a study of the existing codes at present in view of these Design Goals. Based on the study an earnest appeal is made to the forums concerned to ask for more slots for Tamil in UNICODE.*

The study finds and arrives at the following Design Goals:

1. We must have unique Tamil fonts, which could be used in all widely used computer systems throughout the world.
2. The font system should accommodate all the possible variants of Tamil fonts and the few grantha characters. Each one should have a glyph on its own without any kerning.
3. Any attempt in evolving a unique font system should not sacrifice the aesthetic quality of the printed Tamil text.
4. The font system to be devised should not end up with the ultimate goal of promoting DTP possibilities alone. It should provide easy way for data processing by yielding to sorting and searching operations
5. The font system should give adequate place to the creative imagination of the font designer.

The study at the end justifies the finding that a minimum of 256 slots are to be asked for Tamil and any encoding scheme must utilise them to have unique glyph based characters.

Tamil has subjected itself to various and varying factors of influence and has changed at various points of time. No doubt it has changed succumbing to the demands of its developing society and its people. Now it has to face a major force viz. the spread of information technology without which no walk of life could continue to grow. From the time that computers have started setting in the land of Tamils, throughout the world attempts have been made to use Tamil in computers. While English continues to be the language in the operating systems and software packages in the computers even today, Tamil enthusiasts have not attempted to alter it and introduce Tamil. Most of their efforts ended in finding the language in the screen and to use it to have just texts in Tamil. Even capable Tamil technocrats in this field did not think of the laymen in Tamilnadu who have no other language to communicate to the world. They must have been and are carried away by the faith that most of the Tamil population are educated up to a level of understanding certain core words in English which are found necessary to operate a package in the systems. Unfortunately it is not the case. The use of the systems has penetrated to a great extent in the society that the uneducated while attempting to use the packages found it difficult to comprehend and grasp the use of it. This factor is to be borne in mind when one takes up the task of fixing Tamil in UNICODE. The language should be placed in such a way that it must give way for all computing jobs involving data analysis including in the web and not just for the jobs related to Desk Top Publishing alone. This should be the primary Design Goal while placing Tamil in the UNICODE.

So far, lot of work has been carried out in different parts of the world in developing Tamil word processors. The developers are many in the field and hence it has resulted in many non standard-encoding schemes. The enthusiasm and happiness that one achieved in creating and setting Tamil fonts on the monitor and just a text on the paper overshadowed the handicap in such non-standardization.

But of late the proliferation of the Internet usage and the World Wide Web has made the Tamil using community to go in for standardization. And incidentally the intense striving for a standardization of Tamil has got up at a time when the computing world itself is moving towards a UNICODE, which will provide a universal standard for exchange of data and products. Hence our responsibility in recommending and endorsing a standard code for Tamil has become of vital and utmost importance. To achieve this end we must have our Design Goals defined and designed properly targeting all our needs not only at present and also in future.

Here again we are placed in an embarrassing situation. There have already been attempts towards standardization. Each group and individuals have propelled their word processing packages based on different coding systems to the public. Facility for transmission between these packages is provided only within a few. The developers of these coding systems continue to work with the same zeal and interest. With the result now each one would like to have his or the group's coding system get standardized. This tendency must be avoided and future needs alone be the criteria in setting the standardization.

Among all the attempts made, the coding given in TSCII scheme is of much significance. The Tamil using community must stand grateful to the group that has evolved it. An important point to recall is that it was evolved after hectic work for two years by leading Tamil technocrats on the globe. Among most of the information interchange packages (like Murasu Anjal) also follow the same coding system. This system deprives us of lot of flexibility and benefits, which are available in a monolingual fonts system which is prevalent in most of the DTP packages in Tamilnadu.

First of all if an encoding system for Tamil is to be universal in scope it can be achieved in the given context in monolingual font system alone. In a bilingual font system kerning has to be done and it makes the execution slower.

Secondly if one takes into account of the number of users of bilingual fonts and compare them with the number of users who use only monolingual, the second category outnumbers the first. Moreover a driver that switches to the other font can be easily made available. Hence one need not sacrifice the benefits of a monolingual font system for the sake of the existing bilingual systems.

As a third criterion we can think of the future needs where we require a package like MS Office in Tamil. We may need to search or sort out large databases. This can easily be done at a good speed only in a monolingual font system. Though it could also be tackled in a bilingual system it will take lot of time. Anyhow as many desire phonetic encoding could also coexist as a standard just for information exchange.

Most of us think of inputting the text and getting it in print without much quality. While embedding the glyphs through kerning the beauty of the printed text is lost. The printed text does have an aesthetic appeal, especially to the young learners and at any cost it should be preserved. This can be maintained only in the case of a monolingual font system with reference to Tamil.

While designing a font, a font designer should have adequate place for his creative imagination. This can be made possible only if we have place for all the characters in Tamil through monolingual font system. Each possible character should be represented as a separate glyph. Then only it will be easy for a graphic based system to work fast. We are sure that we will not reverse ourselves back to text based system. Hence to accommodate all the Tamil characters along with the needed few grantha characters we have to follow the monolingual font system.

A change, though for more power and facilities, is looked at with askance by those who are reluctant to change the track. Hence nowadays voice of objection is heard from the quarters of the technocrats who were responsible for the early well-knit attempts in Tamil computing. They even recommend to have some adjustment and sacrifice with the language use. What we do today will be

passed to our future generations. Let our successors do not blame us in future for not resorting to a change for the best. First of all why should we give up certain benefits and why should we sacrifice for a few.

Hence the above arguments place before us the following as our Design Goals before us:-

1. We must have unique Tamil fonts, which could be used in all widely used computer systems throughout the world.
2. The font system should accommodate all the possible variants of Tamil fonts and the few grantha characters. Each one should have a glyph on its own without any kerning.
3. Any attempt in evolving a unique font system should not sacrifice the aesthetic quality of the printed Tamil text.
4. The font system to be devised should not end up with the ultimate goal of promoting DTP possibilities alone. It should provide easy way for data processing by yielding to sorting and searching operations.
5. The font system should give adequate place to the creative imagination of the font designer.

These Design Goals naturally lead us to have at least a minimum of 256 locations in the UNICODE system. But unfortunately the Unicode Consortium based on ISCII has assigned only 128 locations to Tamil. This was done without taking into consideration the complexities of character formation in Tamil. ISCII's goal was to provide a computing environment where users without knowing the Indian languages can learn them through transliteration. In the process of achieving this a lot was sacrificed and end goal was also achieved only to the extent of transliterating proper nouns. Moreover there are lot of difficulties in the ISCII standards which are obviously felt in tasks like sorting, indexing and in identifying the characters.

With the above Design Goals and the encoding schemes available, the Chennai Code Version 1.3 satisfies most of the needs enumerated above. But our needs grow and timely revisions and formation of new codes are necessary. Let us get the required locations from the Unicode consortium and work on.

We are safe in the sense that we have time to press forward our request for more locations. Now it should be our endeavour to seek for a minimum of 256 locations and if possible more. We are not greedy in asking for more; in future we may include a few more characters in Tamil due to the spread of knowledge in new vistas or we may modify some of the existing characters. In the light of some of the world languages that have got more than 256 locations let us also ask for more. More locations will have more flexibility in providing a suitable system in this fast changing and growing world. The situation that we meet today is only a transitory one and we can get things done if we construct or reconstruct our theories first. Let us all work together keeping the future needs and generations in mind. In our world nothing is permanent except the change. Let us change for the best.

# தமிழ் நெடுங்கணக்கில் புதியன புகுதல்

## தமிழில் ஒலிபெயர்ப்பு: போதிய வரிவடிவங்கள் இன்மையும் யூனிகோட்(டு ?) கண்ணோட்டத்தில் தீர்வுக்காண முயற்சியும்


*செ.ச. செந்தில்நாதன்*
*துணை ஆசிரியர், இந்தியா டுடே*
*சென்னை*



### பிழிவு

தமிழ் நெடுங்கணக்குக்கான எழுத்துரு குறியீட்டு அமைப்பும் விசைப்பலகை அமைப்பும் தரப்படுத்தப்படும் இந்த தருணத்தில், தமிழ் நெடுங்கணக்கிலேயே சில புதிய வரிவடிவங்களை சேர்க்கவேண்டும் என்று இந்த ஆய்வுரை கோருகிறது. பிறமொழிகளிலிருந்து ஒலிபெயர்க்கும் போது சில ஒலியன்களுக்கான துல்லிய வரிவடிவம் தமிழில் இல்லாததை சுட்டிக்காட்டும் இந்த ஆய்வு அவற்றுக்குரிய புதிய வடிவங்கள் உருவாக்கப்பட்டு அவையும் தரப்படுத்தப்பட வேண்டும் என்கிறது.

g, ch, d, dh, b, x, z ஆகிய (ஆங்கில) ஒலியன்களுக்கு இணையாக, அவற்றுக்கு தற்போது பயன்படுத்தப்படும் எழுத்துகளின் நீட்சியாக கூடுதல் வரிவடிவங்களை உருவாக்க முனையும் இந்த ஆய்வு, தமிழ் நெடுங்கணக்கின் நெகிழ்வுத் தன்மையையும் ஆராய்கிறது. இந்த புதிய எழுத்துக்களின் முக்கியத்துவம் குறித்து தகவலியல் கோட்பாடுகளின் வழியாக சில முடிவுகளை இந்த ஆய்வு முன்வைக்கிறது.


"மாற்றுவதும் திருத்துவதும் யாருக்கும் எதற்கும் இழிவாகவோ குற்றமாகவோ ஆகிவிடாது. மேன்மையடைவும் காலத்தோடு கலந்து செல்லவும் எதையும் மாற்றவும் திருத்தவும் வேண்டும். **மொழி என்பது உலகப் போட்டிப் போராட்டத்துக்கு ஒரு போர்க்கருவி. போர்க்கருவிகள் காலத்துக்கு ஏற்ப மாற்றப்படவேண்டும்"**

– தந்தை பெரியார்.

### ஒலிபெயர்ப்பும் தமிழ் இலக்கணமரபும் – ஒரு மீள்பார்வை

தமிழ் பல காலகட்டங்களில் சமஸ்கிருதம், பாலி, பிராகிருதம், அரபி, லத்தீன், ஆங்கிலம், இந்தி போன்று பல மொழிகளோடு தொடர்புகொள்ள நேரிட்டது. அவற்றுடன் சொல், கட்டமைப்பு ரீதியிலான கொள்வினை கொடுப்பினைகளையும் மேற்கொண்டிருக்கிறது. இந்த வரலாற்று பின்னணியில், ஒலி பெயர்ப்பு பிரச்னையை தமிழ் எப்படி கையாண்டது என்பதை சுருக்கமாக பார்ப்போம். ஒலிபெயர்ப்பு குறித்த தமிழ் மரபை அறியாமல் அதை மாற்றுவதோ மீறுவதோ இயலாதது. முறையற்றது.

வடமொழியிலிருந்து சொற்களை தமிழுக்கு கொண்டுவரும் போது அதை எப்படி எழுத என்பது குறித்து தொல்காப்பியர் காலத்திலேயே ஒரு மரபு இருந்திருக்கிறது:

"வடசொற் கிளவி வட வெழுத்தொரீஇ
எழுத்தொடு புணர்ந்த சொல்லா கும்மே"

என்ற நூற்பாவின் மூலம் தொல்காப்பியர் கூறிய வரம்பின் அடிப்படையில் தற்சமம், தற்பவம் ஆகிய விதிகள் உருவாகின. இந்த இரண்டுவிதிகளின் மூலமும் தமிழில் வடசொல் ஒலிபெயர்க்கப்படுகிற போது வடமொழிக்கான வரிவடிவம் கலக்காமல் பார்த்துக்கொள்ளப்பட்டது. இணையான வரிவடிவங்கள் இருக்கும்போது அப்படியே மாற்றிக்கையாளலாம் என்பது தற்சம முறை (உதாரணம்: குங்குமம்). இல்லாத போது தமிழில் மிகநெருங்கிய, விதிக்கப்பட்ட எழுக்களை பயன்படுத்த வேண்டும் என்பது தற்பவம் (உதாரணம்: பக்ஷம் – பக்கம்). பிற்காலத்தில் தமிழகத்தில சமஸ்கிருதம் மேலாதிக்கம் வகித்த போதும் மணிப்பிரவாள நடை கோலச்சியபோதும் கூட பெரும்பான்மையான மக்கள் இந்த விதிகளையே பின்பற்றினார்கள். இந்த மரபே பிறகு அரபு, லத்தீன், பிற ஐரோப்பிய மொழிகளிலிருந்து ஒலிபெயர்க்கப்பட்டபோதும் பின்பற்றப்பட்டது. உதாரணம்: செய்யது, முகம்மது, கிறித்து, பிரெஞ்சு, இங்கிலாந்து, ஒல்லாந்தர் போன்ற பெயர்ச் சொற்கள்.

மணிப்பரவாளம் மூலமாக தமிழ் உரைநடையில் பல கிரந்த எழுத்துக்கள் புகுந்தன. ஆனால் அவற்றில், இன்று ஜ, ஸ, ஷ, ஹ, ஆகிய கிரந்த எழுத்துக்கள் மட்டும் பெரும்பான்மையானவர்களால் ஏற்றுக்கொள்ளப்பட்டிருக்கிறது. இந்த நான்கு கிரந்த எழுத்துக்களும் தமிழில் நீடிப்பதற்கு காரணம், இப்போது இவ்வெழுத்துக்கள் வடமொழி சொற்களை எழுதுவதற்காக அல்லாமல் ஆங்கில ஒலிபெயர்ப்புகளை எழுதுவதற்காகவே மிகவும் பயன்படுத்தப்படுகின்றன என்பதே. ஹிமாச்சல் பிரதேஷை இமாச்சலப்பிரதேசம் என்று எழுதுவதை சங்கடமின்றி ஏற்றுக்கொள்கிற இன்றைய பொதுவழக்கு ஜப்பானை சப்பான் என்றோ ஆஸ்திரேலியாவை ஆசுத்திரேலியா என்றோ எழுதுவதை ஏற்றுக்கொள்ளாததன் ரகசியமும் இதுவே. இந்த எழுத்துக்களின் தேவை அதிகரித்ததால் அவை நிலைத்துவிட்டன.

ஆனால், சமஸ்கிருத நெடுங்கணக்கோடு நெருங்கிய தொடர்பிருந்தும் பிற திராவிட மொழிகள் அவற்றை அப்படியே ஏற்றுக்கொண்ட போதும், குறைந்த பட்சம் க, ச, ட, த, ப ஆகிய வல்லினங்களுக்கான ஹகரம் ஏறிய, அதிர்வுடைய, அதிர்வும் ஹகரமும் ஏறிய kha, ga, gha முதலான (வர்க்க) எழுத்துக்களை தமிழ் ஏன் ஏற்றுக்கொள்ளவில்லை என்ற கேள்வி எழுப்பப்படலாம். இதற்கான விடைகளை சொல்வது இங்கு நம் நோக்கமில்லை என்றாலும் நமக்குத் தேவையான பிரச்னையோடு இதற்குள்ள தொடர்பை மட்டும் விளக்கமுறப்படுவோம். உதாரணமாக, தமிழ் க, ka, ga, ha ஆகிய மாற்றொலிகளை (allophones) கொண்டிருக்கிறது. எனவே 'க' வர்க்க எழுத்துக்களில் ga ஒலிக்குறிப்பை தமிழ் 'க' ஓரளவுக்கு ஈடுசெய்து வந்தது. மேலும் ஹகரமேறிய aspirated ஒலியன்கள் தமிழ் இயல்புக்கு மாறானது என்று விடப்பட்டன. அத்துடன் வடமொழி சொல் வரவு தொடர்பாக நிலவிவந்த தற்சம-தற்பவ விதிகளை மாற்றும் அல்லது கைவிடும் தேவை எழவில்லை. எழுந்திருந்தால் புதிய வரிவடிவங்கள் முன்பே வந்திருக்கும். உதாரணமாக, F என்கிற ஆங்கில ஒலி யினை குறிப்பிட்ட ஃப் என்ற புதிய கூட்டெழுத்து உருவாக்கிக் கொள்ளப்பட்டதைப் போல.

இந்த ஒலிபெயர்ப்பு வரலாறு இரண்டு முக்கியமான உண்மைகளைச் சுட்டிக்காட்டுகிறது:

- முதலாவதாக, தேவை ஏற்பட்டபோது – அதன் நிர்பந்தம் அழுத்தியபோது – தமிழ் அயல்மொழி ஒலியன்களுக்காக புதிய வரிவடிவங்களை ஏற்றது.
- இரண்டாவதாக, இந்த ஏற்பும் ஐரோப்பிய மொழிகளிலிருந்து ஒலிபெயர்க்க வேண்டிய நிர்பந்தம் வந்தபோதே பரவலாக ஏற்கப்பட்டது. நீடிக்ககாரணமானது.

எனினும் இந்த மாற்றங்கள் போதுமானவை அல்ல. க, ட, த, ப ஆகிய நான்கு முக்கிய வல்லினங்களும் சொல்லுக்கு நடுவில் மெல்லினத்தை தொடர்ந்து வரும்போது ga, da, dha, ba ஒலிகளை குறிக்கின்றன. எனவேதான், London என்கிற சொல் ஒலிபெயர்க்கப்பட்ட போது இயல்பாகவே துல்லியமாக லண்டன் என்று அமைந்தது. என்றாலும் இதே ஒலிகள் ஒலி பெயர்க்கப்பட்ட சொல்லுக்கு முதலிலும் இறுதியில் மெய்யெழுத்தாக தனித்தும் வரும்போது துல்லியமாக அந்த ஒலிகளை சுட்டப்பயன்படுத்த முடிவதில்லை. (உதாரணத்துக்கு Goliath, CAD) இதற்கான தீர்வு, இந்த ஏழு ஒலியன்களுக்கு சொல்லுக்கு முதலிலும் இடையிலும் இறுதியிலும் எந்த இடத்தில் வந்தாலும் ஒரேமாதிரி ஒலிப்பைத் தருகிற வரிவடிவங்களை உருவாக்குவதே ஆகும். இதற்கான தீரிவை நோக்கிய முயற்சியாக சில ஆலோசனைகளை விவாதத்துக்கு வைக்கவிரும்புகிறேன்.

## நிலவும் வரிவடிவங்களின் நீட்சியாக புதிய வரிவடிவங்களை வருவித்தல்

உலகத்தின் எல்லா மொழிகளிலும் உள்ள எல்லா ஒலிக்குறிப்புகளுக்கும் தமிழில் எழுத்துக்களை உருவாக்கித் தீர்த்துவிடுவது என்பதல்ல இந்த ஆய்வுரையின் நோக்கம். ஆங்கிலத்தில் உள்ள எல்லா ஒலியன்களுக்கு தமிழில் இணைகளை உருவாக்குவதும் கூட இதன் நோக்கமல்ல. பன்னாட்டளவில் பொதுவாக ஏற்கப்படுகிற, ஒலியியல் ரீதியாக தரப்படுத்தும் செய்யப்பட்ட சொற்களை ஆங்கிலம் வழி ஒலிபெயர்க்கும்போது சந்திக்கிற பெரிய பிரச்னையை தவிர்க்கவும் ஒலிபெயர்ப்பின் மூலமாக தகவல் சிதைவதற்கு இடம் தரக்கூடாது என்பதற்காகவும் மட்டுமே இந்த மாற்று ஏற்பாடு வேண்டும் என்று இந்த ஆய்வுரை வலியுறுத்துகிறது.

அதற்கான தீர்வை நோக்கிய ஆலோசனைகள் என்ன ?

- க், ட், த், ப் ஆகிய நான்கு அதிர்வற்ற வல்லினங்களின் அதிர்வுள்ள மாற்றொலிகளான g, d, dh, b ஆகியவற்றுக்கு தனியாக நான்கு புதிய எழுத்துக்களை உருவாக்கி சேர்த்துக் கொள்ளுதல்.

- கிரந்த எழுத்துக்களின் உதவியோடு ச்-இன் மாற்றொலிகளான ஜ்-உம் ஸ்-உம் ஏற்கனவே ஏற்கப்பட்டுள்ளன. ஆனால், ch என்கிற ஒலியனுக்கான தமிழ் வடிவம் உருவாக்கப்பட வேண்டும்.
- x, z ஆகிய ஒலியன்களுக்கும் வரிவடிவம் காணப்பட வேண்டும்.

இதற்கு தேவையான வரிவடிவங்களை எப்படி உருவாக்குவது என்ற கேள்விக்கான பதிலை தமிழில் ஜ, ஸ, ஷ, ஹ, ஃப ஆகியவை ஏற்கப்பட்டதின் பின்னணியிலிருந்து வருவிக்கலாம். முதல் நான்கு எழுத்துக்களும் தமிழ் வரிவடிவத்தோடு ஒட்டிய, தமிழ் நாட்டில் வடமொழியை எழுதுவதற்காக பயன்படுத்தப்பட்ட, தமிழ் எழுத்துக்களோடு வடிவ ஒற்றுமைமிகவும் கொண்ட கிரந்த எழுத்துக்களிலிருந்து எடுக்கப்பட்டன என்பதால் அதில் அந்நியத்தன்மையும் வரிவடிவ வித்தியாசமும் இல்லாமல் போய்விட்டன. ஃப் என்பதும் நடைமுறையில் உள்ள இரு தமிழ் எழுத்துக்களின் கூட்டுதான்.

இந்த அடிப்படையில் பார்க்கையில், புதிய வரிவடிவங்களை உருவாக்குவது என்பதற்கு நடப்பில் உள்ள வரிவடிவங்களின் நீட்சியாக, சில மாற்றங்களை செய்து புதிய வடிவங்களாக அவற்றை அறிமுகப்படுத்துவது என்று அர்த்தப்படுத்திக்கொண்டால் அது சரியான வழியாக இருக்கும் என்று தோன்றுகிறது.

கீழ்கண்ட ஆலோசனைகளை யோசிக்கலாம்:

'\' அல்லது '\' அல்லது '.' என்கிற குறியீடு ஒன்றை வித்தியாசப்படுத்தும் குறியீடாக பயன்படுத்தலாம். உதாரணத்துக்கு, 'க' என்ற வரிவடிவத்துடன் இந்த மூன்றில் ஏதோ ஒரு குறியீட்டை இணைத்து 'க' அல்லது 'க' அல்லது 'க' போன்ற ஒரு வடிவத்தை உருவாக்கி அதை ga என்று உச்சரிக்க வேண்டும் என்று விதிக்கலாம். அதே சமயம் இந்த புதிய குறியீடுகள் ஒலிபெயர்ப்புக்கு மட்டுமே என்று வரம்பிடவேண்டும். இந்த உத்தேச வடிவ அமைப்பில் எடுத்துக்கொள்ளப்பட்ட ஏழு ஒலியன்களுக்கான தமிழ் வடிவங்களை கீழே அட்டவணை இட்டிருக்கிறேன்.

| எழுத்து | நடப்பு ஒலி | கூடுதல் ஒலி | மூன்று உத்தேச வடிவங்கள் | | |
|---|---|---|---|---|---|
| க் | k | g | க் | க் | க் |
| ச் | s/c | ch | ச் | ச் | ச் |
| ட் | t | d | ட் | ட் | ட் |
| த் | th | dh | த் | த் | த் |
| ப் | p | b | ப் | ப் | ப் |
| ஸ் | (s)s | x | ஸ் | ஸ் | ஸ் |
| ஜ் | j | z | ஜ் | ஜ் | ஜ் |

இந்த உத்தேச வரிவடிவங்களின் முக்கிய அம்சங்கள்:

1. இந்த வரிவடிவங்கள் தற்போது மாற்றாக பயன்படுத்தப்படும் வரிவடிவங்களின் நீட்சியே. எனவே புதிய வடிவங்கள் தன் தாய் வடிவத்தின் ஒலியின் மாற்றொலியே என்பதை எளிதில் நிரூபிக்கும். எல்லா புதிய வடிவங்களுக்கும் ஒரே வகையான புதுக் குறியீட்டை அளிப்பதன் மூலமாக அவை ஒரே இனத்தைச் சேர்ந்தவை என்பதையும் காட்ட முடியும் (உதாரணமாக, குறியீடு இருந்தால் அது அதிர்வொலி என்பது போல)

2. புதிதாக தரப்படுத்தபடவுள்ள எழுத்துரு குறியீட்டுமுறை (charecter encoding), விசைப்பலகை இட ஒதுக்கீடு (keyboard layout) ஆகியவையும் இந்த மாற்று எழுத்துக்களை ஏற்பது எளிது. கூடுதல் வடிவங்களான இந்த ஏழையும் குறியீட்டுப்பட்டியலில் கூடுதல் எழுத்துக்களாக ஏற்கமுடியும். விசைப்பலகையைப் பொறுத்தவரை, கூடுதல் குறியீட்டுக்காக ஒரேயொரு விசையை ஒதுக்குவதன் மூலம் பிரச்சனையை தீர்த்துவிடலாம். ஆக, ஒரு சின்னஞ்சிறு குறியீட்டைத் தீர்ப்பதன் மூலமே ஒலிபெயர்ப்பு சவாலை முறியடிக்கமுடியும் என்று தெளிவாக தெரிகிறது.

இங்கு இதே பிரச்னை சம்பந்தப்பட்ட ஒரு முக்கிய முன்னுதாரணத்தை எடுத்துக்காட்டுவது நல்லது. மேலும் அது யூனிகோட் சம்பந்தப்பட்ட ஒன்று என்பதால் தற்போதைய விவாதத்துக்கு அது முக்கியத்துவத்தையும் அளிக்கும்.

## புதிய எழுத்துக்களை சேர்ப்பதற்கு யூனிகோட் தரும் வாய்ப்பு – தேவநாகரி முன்னுதாரம்: ஒரு பார்வை வீச்சில்

The Unicode Consortium வெளியிட்டுள்ள The Unicode Standard, Version 2.0 புதிய எழுத்துக்களை சேர்க்கவிரும்புவர்களுக்கு ஆறுதலைத் தருகிறது. இந்திய மொழிகளைப் பொறுத்தவரை, ISCII – II வையே அது மூல தரப்படுத்தமாக எடுத்துக் கொண்டிருக்கிறது. இந்த தரப்படுத்தம் தமிழைப் பொறுத்தவரை முழுமையானதோ முறையானதோ அல்ல என்கிற குற்றச்சாட்டை மறுக்கமுடியாது என்பதை ஏற்றுக்கொள்ளும் இந்த ஆய்வுரை வேறொரு கோணத்திலிருந்து அதன் அம்சங்களை சிலவற்றை ஒலிபெயர்ப்பு குறித்த ஆய்வுக்கு பயன்படுத்திக்கொள்வதோடு நிறுத்திக் கொள்கிறது.

எல்லா இந்திய மொழிகளுக்கான குறியீட்டுப்பட்டியல்களிலும் சில இடங்களை எதிர்காலத்தில் சேர்க்கப்படக்கூடிய வரிவடிவங்களுக்காக என்று காலியாக ஒதுக்கி வைத்திருக்கிறது. எடுத்துக்காட்டாக, தமிழுக்கென ஒதுக்கப்பட்ட 0B80 – 0BF2 ஆகியவற்றுக்கிடையிலான 115 இடங்களில் 54 இடங்கள் காலியாக <reserved> இடமாக வைத்திருக்கிறது. இந்த இடங்களில் கூடுதல் வடிவங்களை முறைப்படி சேர்த்துக்கொள்ளலாம். உதாரணமாக, கன்னடம் f க்கான புதிய வரிவடித்தை அவ்வாறே சேர்த்துக் கொண்டிருக்கிறது.

தேவநாகரிக்கான குறியீட்டுப் பட்டியலை ஆராயும்போது, ஒரு முக்கிய மாற்றம் தென்படுகிறது. Various Signs என்கிற இனத்தில் '.' என்கிற ஒரு கூடுதல் வரிவடிவம் சேர்க்கப்பட்டு, அதற்கான காரணமாக for extending the alphabet to new letters எனவும் குறிப்பிடப்பட்டிருக்கிறது. இதே முறை குஜராத்தி, ஒரியா, வங்காளம் ஆகிய மொழிகளின் குறியீடுகளிலும் பின்பற்றப்பட்டிருக்கிறது.

இந்த விரிவாக்கமுறையை தேவநாகரி எப்படி பயன்படுத்தியிருக்கிறது என்பதை அதே பட்டியலில் அது சேர்த்துக்கொண்டுள்ள சில புதிய எழுத்துக்களைக் கொண்டு விளக்கமுடியும்.

திராவிட மொழிகளிலிருந்தும் குறிப்பாக தமிழில் இருந்தும் சில பிரத்யேக ஒலிகளை ஒலிபெயர்த்து எழுதுவதற்காக நிலவும் வரிவடிவங்கள் சிலவற்றின் கீழே புள்ளி ஒன்றை வைத்து அதை புதிய வரிவடிவமாக நீட்டித்துக்கொள்ளும் உத்தியை இந்த முறை பின்பற்றுகிறது. உதாரணம்: தேவநாகரியின் ரகர எழுத்துக்கு கீழே ஒரு புள்ளிவைத்து அதை தமிழ் வல்லின 'ற'வை ஒலிபெயர்ப்பதற்கான வரிவடிவம் என்று ISCII – II வரையறுத்திருக்கிறது. பட்டியலிலிருந்து:

**0931** DEVENAGARI LETTER LLLA (*for transcrbing Tamil alveolar r)

இதுபோலவே for transcribing Tamil alveolar n, for transcribing Tamil l என்று குறிப்பிட்டே மூன்று வரிவடிவங்களை சேர்த்துக் கொண்டிருக்கிறது. மேலும் திராவிட மொழிகளின் உயிர்க் குறில்களான எ, ஒ ஆகியவற்றை ஒலிபெயர்க்க இரு கூடுதல் வரிவடிவங்களையும் அது சேர்த்துக்கொண்டிருக்கிறது.

தேவநாகரிக்கான யூனிகோட் ஏற்கப்பட்டுள்ள இந்த மாற்றங்களுக்கு பின்னுள்ள அணுகுமுறை பாராட்டத்தக்கது. எளிதில் மற்றமொழிகளிலும் பின்பற்றத்தக்கது. இதே அடிப்படையில் தமிழ் யூனிகோட் குறியீட்டிலும் ஒரு கூடுதல் குறியீட்டை சேர்த்து புதிய வரிவடிங்களை சேர்க்கவழிசெய்யலாம்.

## முடிவுரையாக

மீண்டும் ஒரு எடுத்துக்காட்டை சுட்டிக்காட்டலாம் என்று தோன்றுகிறது. பல்லாயிரக்கணக்கான கருத்தெழுத்துக்களையும் (ideographs) ஹிரகானா, கடகானா என்று இரு வேறுபட்ட ஒலிப்பியல் நெடுங்கணக்கையும் ஒருசேர பின்பற்றும் ஜப்பானிய மொழியும் கூட, அந்நிய மொழிகளிலிருந்து ஒலிபெயர்க்க நேரிடும்போது துல்லியமாக செய்யவேண்டும் என்பதற்காக கடகானா எழுத்துக்கள் சிலவற்றை நீட்டித்து சுமார் 20 புதிய வரிவடிவங்களை உருவாக்கிக்கொண்டது. ஜப்பானிய மொழியின் அல்லது இந்தியின் அணுகுமுறை இந்த வகையில் முற்போக்கானது. காலத்தை ஒட்டியது. இது வளரும் மொழிகளுக்கான அடையாளம்.

இந்த விதி தமிழுக்கும் பொருந்தும். நம்மிடமே அதற்கு முன்னுதாரணம் இருப்பதையும் முன்பே பார்த்தோம். இப்போது சற்று கூடுதலான வரிவடிவ தேவைகளை நிரப்பவேண்டிய தருணத்தில் உள்ளோம். தற்போதுள்ள தமிழ்நெடுங்கணக்கில் புதிய வரவுகளாக இவற்றை இந்த ஆய்வுரை முன்மொழிந்தாலும் விருப்பத்தின் அடிப்படையில் இவற்றை பயன்படுத்துவதற்கான ஒரு வாய்ப்பாவது அளிக்கப்படலாம்.

இதுகுறித்து விரைந்து விவாதித்து துணிந்து முடிவெடுத்து அவற்றையும் தரப்படுத்திவிட்டால் அது தமிழுக்கு கூடுதல் பலத்தை தரும் என்பதை தொழில்முறை மொழிபெயர்ப்பாளன் என்கிற என் சொந்த அனுபத்திலிருந்து நிச்சயம் அடித்துச் சொல்லமுடியும்.

## Selective Bibliography


1. The Unicode Standard, Version 2. The Unicode Consortium
2. ISCII and Tamil – A Perspective. By N. Anbarasan, paper presented in TamilNet 97 Conference.
3. பேச்சொலியியல், சி.சுப்பிரமணியன், நாட்டார் வழக்காற்றியல் ஆய்வு மையம், பாளையங்கோட்டை, 1998.
4. பொது மொழியியல், டாக்டர். பொற்கோ, பூம்பொழில், 1997.
5. அறிவியல் தமிழ், டாக்டர். வா.செ. குழந்தைசாமி, பாரதி பதிப்பகம், 1985.

*SECTION - III*

# Use of Tamil in Multimedia Internet and Computing

# A TAMIL SPEECH SYNTHESIS SYSTEM

*Ramakrishnan, A. G., Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012.*
*and*
*V. Karthigeyan, Ncore Systems, Bangalore*

**Abstract**

*This paper describes the development of a limited vocabulary, preliminary speech synthesis system for Tamil, incorporating some rules representing the knowledge sources. A speech synthesis system has many applications, such as natural language interface for computers, multimedia education packages in Tamil, automatic telephone-based enquiry systems in Tamil Nadu Government organisations, audio on-line help in Tamil based IT software, Virtual teacher on CD ROM/ Internet, computer based Tamil teaching, automatic document reading machines in Tamil for the blind, speech output in Government Online projects (such as that proposed with World Tel), and intelligent toys/ multimedia aids for children. Speech synthesis involves synthesizing intelligible and natural-like speech from text in coded form (ISCII or unicode). This involves two major phases (i) the text analysis phase and (ii) the speech synthesis phase. The text analysis phase parses the input text into a sequence of basic units of speech. The speech synthesis phase involves the concatenation of the parameters of these units and synthesis, after the application of both segmental and suprasegmental rules for naturalness. Since Tamil is phonetic in nature, characters are chosen as the appropriate basic units. Acoustic parameters such as linear prediction coefficients, formant frequencies, bandwidths, pitch and gain are prestored for the basic speech sound units corresponding to the orthographic characters of Tamil. The parameters are concatenated based on the input text for synthesizing speech. In order for the synthesized speech to appear natural, we need to adequately model the various features of natural speech, such as coarticulation, intonation and duration. These knowledge sources are stored in the form of rules. Coarticulation rules specify the pattern of joining the basic units. Intonation rules specify the overall pitch contour for the utterance. Duration rules modify the duration of the basic units based on the linguistic context in which they occur.*

## Introduction

Speech synthesis involves synthesizing, intelligible and natural-like speech from unrestricted text of a language. Progress in speech synthesis has been made possible by advances in linguistic theory, acoustic-phonetic charecterisation of sound patterns, mathematical modelling of speech production, structured programming and computer hardware design. Potential Applications of a speech synthesis system in Tamil include

1. Natural language interface for computers
2. Self-learning multimedia education packages in Tamil
3. Automatic telephone-based enquiry systems in all Tamil Nadu Government organisations
4. Audio on-line help in (futuristic, but positively feasible) Tamil based IT software
5. Virtual teacher on CD ROM/ Internet (for distance learning/ Open Universities)
6. Computer based Tamil teaching (speaking books, multimedia dictionaries)
7. Text to speech in Tamil and voice mail
8. Automatic document reading machines in Tamil for the blind

9. Learning IT in Tamil

10. Speech Output in Government Online projects (such as that proposed with World Tel)

11. Intelligent teaching toys/ multimedia aids for children.

Developing a complete speech synthesis system requires proper insight into the human speech production mechanism and knowledge of the perceptual significance of various acoustic parameters of the speech model. The major issues in the design of a speech synthesis system are the following:

? Choice of an appropriate speech synthesis model.
? Collection of data required for segmental synthesis.
? Acquisition, representation and incorporation of various knowledge sources that account for the naturalness of speech.

The common approaches followed for speech synthesis are :

? Concatenation method.
? Synthesis by rule method.

The concatenation method involves collecting and storing the basic units of speech [O' Shaugnessy 1984, Lukaszewicz et al 1987]. The basic units in the text to be synthesized are identified and the stored basic units are concatenated to get the speech output. In synthesis by rule method, a set of rules operate on the given text to produce a parameter sequence which is then synthesized. The rules are for the generation of basic speech units (acoustic -phonetic knowledge) and for the naturalness of the synthesized speech (Allen 1976).

## Materials and Methods

Characters are more compact than phonemes, but their sounds vary with the place of occurrence in a word, especially for non-phonetic languages. Whereas for phonetic languages, each character has a unique script and pronunciation. Since Tamil is phonetic in nature, character could be an appropriate choice as basic unit, which we have choosen. A character in Tamil can be any one of (i) consonant-vowel sequence (CV) (ii) consonant alone (C) (iii) vowel alone (V). Cluster characters such as CCV and CCCV etc can be generated by combining appropriate Cs and a CV. The other advantages in choosing the character as the basic unit are

? The number of characters are not so large (around 250 in Tamil).
? Consonant to vowel transitions are preserved in the characters itself.
? characters have a more natural pronunciation than the phonemes, diphones, etc.

## Model of speech production

Speech sounds are produced as a result of acoustic excitation of the human vocal tract. When the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords, the sounds produced are termed as voiced sounds. All the vowels come under this category. When the excitation is produced by air passing turbulently through constrictions on the vocal tract, the sounds produced are called unvoiced. The consonants like \pa\ & \sa\ come under this category. The speech production model has been developed with an assumption that the variations in the vocal tract shape with time can be approximated with sufficient accuracy by a succession of stationary shapes. These set of stationary shapes of the vocal tract are modelled as discrete time varying digital filters. With this basic idea of speech production model, the various acoustic parameters can be understood well. The acoustic parameters and their relation to the human speech production model are described below.

? Pitch : It is the fundamental frequency, $F_0$ of the vocal cords. The typical values are 50-200 Hz for men, 150-300 Hz for women and 200-400 Hz for children.

? Formants : These are the resonant frequencies of the vocal tract. Their values depend on what is uttered. Only the first three formant frequencies are perceptually significant and they lie below 3000 Hz.

? Bandwidths : Bandwidths determine the concentration and distribution of energy near the formant frequencies in the spectra of the signal.

? LPC : These are the linear predictive coefficients, i.e, filter coefficients of the all-pole model of the vocal tract.

? Gain : Energy of vibration of the vocal cord.

## Construction of the basic unit database

The basic unit data base has to be developed with great care as the quality of the synthetic speech predominantly depends on the quality of the basic units. Certain guidelines need to be followed in developing the database. They are as follows.

? The basic speech unit has to be embedded in some carrier word which needs to be meaningless in order to avoid any prosodic bias.

? The carrier word should be so chosen that the basic unit of interest is easily separable from it. The motivation behind extracting the basic units from some carrier word rather than uttering it in isolation is that, we can use the duration of the basic unit extracted from the carrier word as the default length for synthesis. The carrier words recorded are digitized at a sampling rate of 10 KHz. Waveforms of the basic speech units are carefully separated from the carrier words using an interactive audio editor package, Goldwave.

## Pitch extraction

We have used the simplified inverse filtering technique (SIFT)} for pitch extraction (Markel 1972). The speech waveform s(t) is first filtered by a lowpass filter with a cut-off at 0.8 KHz. After sampling the filter output at 2 kHz, the first five terms of the short term autocorrelation sequence are calculated for an appropriate input length (corresponding to 32 msec) of data. Then a set of linear equations are solved for inverse filter coefficients $a_i$. Knowing $a_i$, the inverse filter output $y_n$ can be calculated. The output autocorrelation sequence $r_n$ is then calculated as the autocorrelation sequence of $y_n$. After $r_n$ is obtained, the largest peak within specified limits is found. If the peak is above a threshold (0.4), then that frame is classified as voiced and the reciprocal of the location of the peak is the pitch, $F_0$. If the peak is less than the threshold, the corresponding frame is classified as unvoiced, with $F_0 = 0$. The advantage of the SIFT algorithm over other algorithms is that, the voiced and the unvoiced regions are satisfactorily classified with the same threshold for all the frames.

## Linear prediction coefficients and Gain computation

We have used the Levison-Durbin's recursion algorithm to extract the linear prediction coefficients and gain for each frame of the data. Linear prediction synthesis There are two parametric models of speech synthesis. They are
? The linear prediction synthesis.
? The formant synthesis.

The linear prediction (LP) method (Makhoul, 1975) is based on the speech production model, where the vocal tract is modelled as an all-pole digital filter of order 'p' and gain, G. Here, every speech sample is computed as the linear combination of a certain no of previous speech samples and a sample of the excitation signal. The advantages of the LP model are the following: ? It can represent fairly well most of the speech sounds except nasals and voiced fricatives. ? Efficient methods for automatic extraction of linear prediction coefficients are available. The disadvantages are the following: ? It cannot model zeros in the speech spectrum and hence it does not model speech sounds like nasals satisfactorily. ? It is very rigid against spectral level modification. Because of these disadvantages, adjustments of peak frequencies (formants) in the spectrum are not possible. The second disadvantage is very severe making it impossible to incorporate the knowledge due to coarticulation which modify the spectral peaks contextually. The formant synthesis requires the estimation of bandwidths from the spectrogram of the signal. The spectrogram required for the accurate measurement of bandwidth should be of very good resolution. Since we do not have the necessary equipment to get a very good resolution spectrogram, we could not use the formant synthesis.

## Excitation source

The quality of the synthetic speech generated by these parametric models depends on the excitation signal used as input. The excitation signal should be such that its spectrum is white. This is because, it is easier to shape the white spectrum according to a model, compared to a colored spectrum. Based on the type of the sound to be synthesized, the excitation signal should be either a white noise sequence (for unvoiced sounds) or periodic pulses a pitch period apart (for voiced sounds). We have used the Fant's excitation model for excitation signal generation (Childers, 19). In this model, the energy is distributed over one pitch period. Hence the synthesized speech obtained using this type of excitation is significantly better compared to the speech generated using impulse excitation. The parameters of the model are T1 =50% of pitch period and T2 =12% of pitch period.

## Knowledge sources for naturalness

Mere concatenation of the signals corresponding to the basic units of speech does not produce intelligible and natural sounding speech. To produce speech from a given text, human beings use several knowledge sources such as phonetics, phonology, morphology, syntax, semantics and pragmatics (V.R. Ramachandran, 1991). It is necessary to incorporate these knowledge sources in a suitable form in the system to accomplish the same task. Human speech is characterised by segmental and suprasegmental features which collectively contribute to the naturalness of speech. A segment refers to some small chosen unit of speech (eg phonemes, syllables etc). Segmental features refer to those which decide the phonetic quality of the segment. Suprasegmental or prosodic features have their domain extended over more than one segment. The suprasegmental features are influenced by factors such as phonetic and syntactic context, semantics and emotional state of the speaker. In continuous speech, the segmental features are subjected to changes decided by phonetic context. These changes are caused by the coarticulation effect and give rise to certain joining patterns between adjacent speech units. One needs to acquire the knowledge pertaining to these features from natural speech. Their incorporation into the system makes the synthesized speech sound intelligible and natural. Some of the knowledge sources identified here are coarticulation, intonation and duration.

## Intonation in Tamil

With the knowledge of the general declining tendency of the pitch contour of declarative sentences (A.S.Madhukumar et al,1991), we have developed a model to synthesize intonation pattern for simple declarative sentences. The model is developed based on the observations from a set of 25 declarative sentences read out by an adult, male, native speaker of Tamil. The model assumes that ? F_0 for any simple declarative sentence fluctuates between two abstract lines, a base line and a top line. ? The top line is determined by the value of the first peak of the first word and the last peak of the last word of the sentence.

? The base line is determined by the valley following the first peak of the first word and the valley of the final word of the sentence.

? The intonation contour is charectarized by the local falls and rises of F_0 based on the stress on the words of the sentence.

The F_0 values of the valleys and the peaks in each word are computed from the corresponding values taken from the base line and the top line respectively. The intonation knowledge modifies the default pitch values of the basic units. The nature of the underlying sentence (declarative, interrogative, yes/no type, etc.) is determined and based on that, the corresponding intonation is provided to the pitch contour, prior to synthesizing.

## Incorporation of duration knowledge

Studies have been made to determine the change of duration of the basic units in different contexts (B.Yegnanarayana, 1994). From these studies, rules are formulated to modify the duration of the basic units during synthesis, depending on the context in the input text. Some of the rules are the following.

? Positional effect: The duration of a character at the beginning of a word is increased by 10% and that at the final position, is increased by 30%.

? Prepausal lengthening effect: The duration of a character appearing before a pause is increased. If the pause is due to a punctuation like comma, then the final character of the word preceding the pause is increased by 35\%. If the pause is due to sentence ending, then the increase will be by 35\%. If the pause is due to any other syntactic boundary, then the duration of the character will be increased by 30%.

The base duration, D_0, of the basic unit is taken from the carrier word. Since the carrier word is spoken in isolation, the actual base duration in continuous speech is about 70% of the corresponding D_0. Therefore, the actual base duration, D for each basic unit is taken to be 70% of D_0. Now, depending upon the context of each basic unit, various rules may be applied. Each of these modifies the base duration, D for that particular unit using the expression D_{final} =D +alpha*D/100 where alpha is the percentage value specified in each rule. The basic units are synthesized for D_{final} duration. The duration knowledge does not modify any of the other parameters used for synthesis.

## Conclusion

The present speech synthesis system for TtS conversion in Tamil is successful in generating all vowels, all CV units except those that have the consonants \r/, \v/ & \p/. The consonants have not been synthesized properly owing to their very short duration. Words without consonants have been synthesized with a reasonable amount of intelligibility using the present speech synthesis system. Attempts have also been made to synthesize some declarative sentences. Proper pitch contour has been provided to the declarative sentences synthesized, in order to give proper intonation.

The main contributions of this work are the following:

1. The basic units of speech that are suitable for speech synthesis in the context of TtS conversion are collected from natural speech.
2. Some knowledge sources that account for the naturalness of speech are identified and acquired.
3. A representation scheme is proposed for the basic speech units which is flexible enough to allow incorporation of the knowledge sources.
4. Demonstration of a limited vocabulary speech synthesis system for a TtS conversion system in Tamil, incorporating some of the rules that represent the knowledge sources.

Text to Speech conversion has been one of the focus areas of research for the speech scientists for over three decades. Currently available TtS conversion systems in English, that are capable of producing highly intelligible speech, are the result of this research. The foci of research have only been English, French, German and Hindi. As far as our knowledge goes, no work has been done in developing a TtS conversion system for Tamil. So, this is the first attempt in this direction.

## References


1. B. Yegnanarayana, "Formant Extraction from linear-prediction phase spectra," J. Acoust. Soc. Am., Vol. 63, No. 5, pp. 1638-1640, May 1978.
2. Hema A. Murthy and B. Yegnanarayana, "Speech processing using group delay functions," Signal Processing, Vol. 22, pp. 259-267, 1991.
3. Dennis H. Klatt, "Review of text-to-speech conversion for English," J. Acoust. Soc. Am., Vol. 82, No. 3, pp. 737-793, May 1987.
4. B. Yegnanarayana, S. Rajendran, V. R. Ramachandran and A. S. Madhukumar, "Significance of knowledge sources for a text-to-speech system for Indian languages," Sadhana, Vol. 19, Part 1, pp. 147-169, Feb. 1994.
5. Eric Moulines and Francis Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," Speech Communication, Vol.9, pp. 453-467, 1990.
6. Francis Charpentier and Eric Moulines, "Text-to-speech algorithms based on FFT synthesis," Proc. ICASSP'88, pp. 667-670, 1988.
7. Bishnu S. Atal and Nancy David, "On synthesing natural-sounding speech by linear prediction," Proc. ICASSP'79, pp. 44-47.

8. S. E. G. Ohman, "Coarticulation in VCV utterances: spectrographic measurements," pp. 151-168, 1965.

9. A. S. Madhukumar, S. Rajendran, and B. Yegnanarayana, "Intonation component of a text-to-speech system for Hindi," Computer Speech and Language, Vol. 7, pp. 283-301, 1993.

10. Andrew Varga and Frank Fallside, "A technique for using multipulse linear predictive speech synthesis in text-to-speech type systems," Proc. ICASSP'87, pp. 586-587.

11. D. G. Childers and H. T. Hu, "Speech synthesis by glottal excited linear prediction," J. Acoust. Soc. Am., Vol. 96, No. 4, pp. 2026-2036, Oct. 1994.

12. N. Yiourgalis and G. Kokkinakis, "A TtS system for the Greek language based on concatenation of formant coded segments," Speech communication, Vol. 19, pp. 21-38, 1996.

13. Jonathan Allen, "Synthesis of speech from unrestricted text," Proc. IEEE, Vol. 64, No. 4, pp. 433-442, Apr. 1976.

14. J. D. Markel and A. H. Gray, "A linear prediction vocoder simulation based upon the autocorrelation method," IEEE Trans. ASSP, Vol. No. pp. 124-134, Apr. 1974.

15. D. H. Klatt, "Software for a cascade/ parallel formant synthesizer," J. Acous. Soc. Am., Vol. 67, No. 3, pp. 971-995, 1980.

16. B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," J. Acous. Soc. Am., Vol. 67, No. 3, pp. 637-655, 1971.

17. J. D. Markel, "Digital inverse filtering - a new tool for formant trajectory estimation," IEEE Trans. Audio, Electroacoust., Vol. 20, No. 2, pp. 129-137, June 1972.

18. Noriko Umeda, "Linguistic rules for Text-to-Speech synthesis," Proc. IEEE, Vol. 64, No. 4, pp. 443-451, Apr. 1976.

19. J. D. Markel, "The SIFT algorithm for fundamental frequency estimation," IEEE Trans. Audio, Electroacoust., Vol. 20, No. 5, pp. 367-377, 1972.

20. J. P. Olive, "Mixed spectral representaion-formants and linear predictive coding, J. Acoust. Soc. Am., Vol. 92, No. 4, pp. 1837-1841, 1992.

# உலகளாவிய மின் வலையில் தமிழ்ப் பாடம்

அரங்க வாசு
துணை இயக்குனர், பென் மொழிகள் மையம்,
பென்சில்வேனியா பல்கலைக் கழகம், பிலடெல்பியா, (USA)

தேமதுரத் தமிழோசை உலகமெலாம் பரவும் வகை செய்தல் வேண்டும்

—மகாகவி பாரதியார்

**கட்டுரைச் சுருக்கம்**

மின்வலையினை பயன்படுத்துவோரின் எண்ணிக்கை அதிகரித்த இக்காலக்கட்டத்தில், இதன் மூலம் தமிழ் மொழிப் பாடங்களைக் கொடுப்பது மிகவும் எளிதான ஒன்றாக இருக்கிறது. குறிப்பாக தமிழ் கற்போரின் பேசும் திறன் மற்றும் எழுதும் திறனை அதிகரிக்கும் வகையில் பாடங்களை அமைத்துத் தருவதில் மின்வலைப் பக்கத்தை உருவாக்குவோர் கவனம் செலுத்தவேண்டும். மேலும், அயல் நாட்டில் தமிழ் கற்போருக்கு பேச்சுத்தமிழ் மற்றும் எழுத்துத் தமிழ் ஆகிய இரண்டு வகைத் தமிழின் வேறுபாட்டை விளக்கவேண்டியிருக்கிறது. இக்கட்டுரை மின்வலையில் கொடுக்கத் தகுந்த தமிழ்ப் பாடவகைகளையும் பென்சில்வேனியாப் பல்கலைக் கழகத்தில் அமைக்கப்பட்டுள்ள தமிழ் மின்வலைப் பக்கத்தில் கொடுக்கப்பட்டுள்ள தமிழ்ப் பாடங்களையும் பற்றி விளக்குகிறது.

## 0. முன்னுரை

பாரதியின் கனவு நனவாக உலகம் முழுதும் தமிழர்கள் வாழ்ந்து வளர்ந்து நமது தமிழ் மொழிக்கும் தமிழ்க் குலத்துக்கும் பெருமை சேர்த்து வருகிறார்கள். தமிழர்கள இல்லாத நாட்டினை விரல் விட்டு எண்ணிவிடுகின்ற அளவுக்கு அவர்கள் திரும்பிப்பார்க்கும் திசையிலெல்லாம் இருக்கிறார்கள். தமிழகத்தில் வாழும் தமிழ்க் குழந்தைகளுக்குத் தமிழ் கற்றுக்கொடுப்பதில் சிரமமேதும் இல்லை. ஆனால் மேலை நாடுகளில் வாழும் தமிழ்க் குழந்தைகளுக்குத் தமிழ் கற்றுத் தருவது என்பது சற்றுக் கடினமான செயலாகும். இதற்கு முக்கிய காரணம் இவர்கள் தமிழ் மொழியை இரண்டாவது மொழியாகக் கற்றுக்கொள்கிற சூழ்நிலைக்கு உள்ளாகிறார்கள். தமிழ் மொழியை இரண்டாம் மொழியாக கற்றுத் தருவதற்கான பாட நூல்களும் வசதிகளும் நிறைய இருக்கின்றன என்று சொல்லிக்கொள்ள முடியாத நிலையில் இருக்கிறோம். உலகலாவிய மின்வலை பெருகிவிட்ட இச்சூழ்நிலையில் தமிழ் மொழியை இரண்டாம் மொழியாகக் கற்றுக்கொடுப்பதற்கான வசதிகள் மிகவும் வளர்ந்திருக்கிறது.

இக்கட்டுரை மின்வலை மூலம் தமிழ் கற்றுக்கொடுப்பதன் வசதிகளையும், பென்சில்வேனியா பல்கலைக் கழகத்தில் அமைக்கப்பட்டுள்ள தமிழ் மின்வலைப் பக்கத்தைப் பற்றியும் விளக்குகிறது[1]

## 1. மொழித் திறனும் மின்வலையின் பயன்பாடும்

மொழி கற்போரின் திறனை இரண்டு வகையாகப் பிரிக்கின்றனர். அவைகள்: 1) உள்நிலை அறிவு (passive knowledge), 2) செயல் நிலை அறிவு (active knowledge) என்பவையாகும். படித்தல் மற்றும் புரிந்துகொள்ளல் (reading and understanding) உள்நிலை அறிவாகும். பேசுதல் மற்றும் எழுதுதல் (speaking and writing) செயல் நிலை அறிவாகும். மேலை நாடுகளில் வளர்ந்து வரும் தமிழ் குழந்தைகளுக்கு உள்நிலை அறிவைப் பெறுவதில் எந்தவித பிரச்சினையும் இருப்பதில்லை. பெற்றோர்கள் மற்றும் உறவினர்கள் வீட்டில் பேசுவதன் மூலம் அவர்களால் தமிழை மிக நன்றாகப் புரிந்துகொள்ள முடிகிறது. சிறிது முயற்சிக்குப் பின் அவர்களால் தமிழ் எழுத்தைப் புரிந்துகொண்டு மிக நன்றாகப் படிக்கவும் முடிகிறது. ஆனால் அவர்களுக்குச் செயல் நிலை அறிவான பேசும் திறனையும் எழுதும் திறனையும் புகட்டுவதில் மிகுந்த சிரத்தை எடுக்க வேண்டியிருக்கிறது. இதற்கு முக்கியக் காரணம் இவர்கள் தாங்கள் வகுப்பில் கற்றுக்கொள்ளும் பாடங்களை பெற்றோர்களிடம் மட்டும்தான் பேசிப் பழக வாய்ப்பிருக்கிறது. தாய், தந்தை இருவரைக் கொண்டு தமிழைச் சரளமாகப் பேசக் கற்றுக்கொள்ள முடியும் என்பது இயலாத ஒன்றாகும். எழுதும் திறனைப் பெறுவதற்கும் அவர்கள் இடைவிடாத முயற்சி செய்ய வேண்டியுள்ளது. இதற்கெல்லாம் இடையூறாக அவர்கள் கற்றுக்கொண்ட முதல் மொழி குறுக்கே நிற்கிறது. அவர்களால் எந்தச் சூழலிலும் முதல் மொழியைக் கொண்டு சிறப்பாக இயங்க முடிகிற நிலையில் தமிழ் மொழி இடைச் செறுகலாகவே அவர்களுக்குத் தோன்றுகிறது.

## 2. மேலை நாடுகளில் தமிழறிவு

தமிழர்கள் உலகம் முழுவதும் பரவிவிட்ட இந்த நிலையில் தமிழ் மொழியும் அந்த அளவிற்கு வளர்ந்து வருகிறதா என்ற வினா எழுப்பின், இல்லை என்றே சொல்லத் தோன்றுகிறது. மேலை நாடுகளில் வாழும் தமிழ்ச் சந்ததியினரின் தமிழறிவு பெருமளவில் குறைந்து காணப்படுகிறது. பெரும்பாலான தமிழ்க் குழந்தைகள் தமிழை இரண்டாவது மொழியாகவே கற்கின்றனர்.

[1] உங்கள் மின்வலையை அலசும் கனிமத்தைக் கொண்டு http://ccat.sas.upenn.edu/plc/tamilweb/ என்ற பக்கத்தைத் தயவுசெய்து நோக்குங்கள்.

அமெரிக்காவைப் பொறுத்த வரையில் யாராவது ஒரு ஏழு வயது தமிழ்ப் பையன் தமிழ் பேசினால், அதைப் பார்க்கிற மற்றத் தமிழ்ப் பையன்களுக்கு அது மிகவும் ஆச்சரியமாக இருக்கிறது. Mom! look! He speaks Tamil! என்று கூறி ஆச்சரியப்படுகிறான். இங்கே வசிக்கிற தமிழ்க் குழந்தைகள் சனிக்கிழமையிலோ ஞாயிற்றுக் கிழமையிலோ பெற்றோர்களின் முயற்சியால் நடத்தப்படும் மரத்தடிப் பள்ளி போன்ற வகுப்புக்களில் விளையாட்டாகத் தமிழ் கற்கின்றனர். அவர்கள் கல்லூரிக்கு வரும் போது ஒரு சில பல்கலைக் கழகங்களில் நடத்தப்படும் Tamil 230, Tamil 231 என்று எண்கள் கொடுக்கப்பட்ட தமிழ் வகுப்புகளில் சேர்ந்து நாளுக்கு ஒரு மணி நேரம் விதம் ஒன்று அல்லது இரண்டு வருடங்களுக்குத் தமிழ் பயின்று பாட்டிக்கும் தாத்தாவுக்கும் தமிழில் கடிதம் எழுதிப் பெருமைப் படுகின்றனர். இவர்களுக்குத் தங்களின் தாயைப் போலவோ தந்தையைப் போலவோ தமிழ் பேசமுடியவில்லையே என்ற ஆதங்கம் இருந்து வருகிறது. அங்கங்கே ஒரு சில மாணவர்கள் தங்களின் இடைவிடாத முயற்சியில் தாங்களாகவே புத்தகத்தையும், கணிப் பொறியையும் பயன்படுத்தி தமிழை நன்றாகப் பயின்று வருகின்றனர். எது எப்படி இருப்பினும் நாமெல்லாம் பருகி மகிழும் தேமதுரத் தேனாகிய இலக்கியத் தமிழை இவர்கள் பருகிக் களிக்க வேண்டுமானால் பெற்றோர்களின் கடும் உழைப்பும் கல்வியாளர்களின் அயராத ஒத்துழைப்பும் மிகவும் அவசியம்.

### 3. பேச்சுத் தமிழும் எழுத்துத் தமிழும்

இரண்டாவது மொழியாகத் தமிழ் மொழியினைக் கற்போருக்குப் பேசவும் எழுதவும் தேவையான அறிவைப் புகட்டுவதே முக்கியக் குறிக்கோளாக அமைகிறது. தமிழ் நாட்டில் தமிழ் படிக்கும் மாணவர்களுக்குப் பேசும் திறனைப் புகட்டவேடியதில்லை ஏனெனில் அது மீன் குஞ்சுக்கு நீந்தக் கற்றுக் கொடுப்பது போலாகும். மற்ற மொழிகள் போல் தமிழ் மொழி இல்லை. நாம் பேசுவது ஒரு மாதிரி எழுதுவது ஒரு மாதிரி. அதோடு, அப்பாவிடம் பேசுவது ஒரு மாதிரி ஆசிரியரிடம் பேசுவது ஒரு மாதிரியான இரட்டை வழக்கு (Diglossia) நிலை. மேலை நாடுகளில் குறிப்பாக அமெரிக்கா, கனடா, ஆஸ்திரேலியா போன்ற நாடுகளில் தமிழ் படிக்கும் தமிழ் மாணவர்களுக்குப் பேச்சுத் தமிழைச் சொல்லிக்கொடுப்பது மிகவும் அவசியமானதாக இருக்கிறது. ஏனெனில் இவர்களுக்கு சிங்கப்பூர், மலேசியா போன்ற நாடுகள் போல் அல்லாது தமிழ் பேசுவதற்கான சூழ்நிலைகள் அவ்வளவாகக் கிடைப்பதில்லை. பென்சில்வேனியாப் பலகலைக் கழகத்தின் தமிழ் வலைப் பக்கம் மூலம் கொடுக்கப் படும்

பேச்சுத் தமிழுக்கான பாடம் ஒன்றை சிங்கப்பூரில் நடந்த கருத்தருங்கில் விளக்கிக் கொண்டிருந்த போது ஏற்பட்ட சம்பவம் ஒன்றை இங்கு குறிப்பிடுகிறேன். பின்வரும் தொடர் போன்று வரும் பேச்சுத் தமிழ் கொண்ட ரேடியோ நாடகத்தின் பகுதிகளை எங்கள் மின்வலையில் கொடுத்திருக்கிறோம்.

" அப்பாப்பா! அந்த ஆளு ஒங்களெ பாத்துட்டாருப்பா. கையெக்கூட ஆட்டுறாருப்பா! பாருப்பா!"

"அடப் பாவி மனுஷா! பாத்திட்டானா இனிமே ஓடினா பிரயோசனம் இல்லெ. உம்.. இன்னெக்கி யாரு மொகத்துலெ முழிச்சேனோ!..."

இம்மாதிரியான தமிழை மாணவர்களுக்குச் சொல்லிக்கொடுப்பதில் தனக்குரிய அதிருப்தியை ஒரு சிங்கப்பூரில் வாழும் தமிழ் ஆசிரியர் வெளிப்படுத்தினார். இந்தப் பகுதியினை நாங்கள் மூன்றாம் நிலையிலான (Advanced Tamil) மாணவர்களுக்குக் கொடுக்கிறோம். இதில் உள்ள "பாவி மனுஷின்", "யார் முகத்தில் முழிச்சேனோ" போன்ற வழக்குத் தொடர்களைக் கற்பிப்பதன் மூலம் தமிழ் மொழியை மட்டுமல்லாது அவர்களுக்கு தமிழர்களின் பழக்கவழக்கங்கள் மற்றும் பண்பாடு ஆகியவற்றையும் கற்பிக்க வசதியாக இருக்கும். இது எந்த வகையிலும் மாணவர்களுக்கு பொருந்தாத ஒன்று என்று வாதிட முடியாது. பெரும்பாலோர் தமிழ் வகுப்பு என்றால் மாணவர்களுக்கு இலக்கியத் தமிழைத் தான் சொல்லிக்கொடுக்கவேண்டும் என்ற தவறான எண்ணத்தைக் கொண்டுள்ளனர். இது தமிழ்நாட்டில் வாழும் குழந்தைகளைப் பொறுத்த வரையில் சரியாக இருக்கலாம் ஆனால் மேலை நாடுகளில் வாழும் குழந்தைகளுக்குப் பொருந்தாத ஒன்றாக இருக்கும். எழுத்துத் தமிழை மட்டுமே இவர்களுக்குக் கற்றுக் கொடுத்தால் சென்னையில் வந்து இறங்கியவுடன் நமது மாணவர்கள்:

இங்கிருந்து வள்ளுவர்கோட்டத்துக்கு எப்படி போவது என்று நீங்கள் தயவுசெய்து எனக்கு வழிசொல்கிறீர்களா?

என்று ரிக்ஷா ஓட்டுனரிடம் கேட்பார்கள்.

இங்கேருந்து வள்ளுவர்கோட்டத்துக்கு எப்படிப் போறதுண்ணு நீங்க கொஞ்சம் எனக்கு வழி சொல்றீங்களா?

என்ற பேச்சுத்தமிழ் வாக்கியங்களை மேலை நாடுகளில் தமிழ் பயிலும் மாணவர்களுக்கு சொல்லிக்கொடுக்கவேண்டியிருக்கிறது. ''படித்துக்கொண்டிருக்கிறேன்'' என்று எழுதவேண்டும். ''படிச்சிக்கிட்ருக்கேன்'' என்று சொல்லவேண்டும்; எடுத்துக்கொள்வேன் என்று எழுதவேண்டும் எடுத்துப்பேன் என்று சொல்லவேண்டும் என்பது போன்ற எழுத்துத் தமிழுக்கும் பேச்சுத்தமிழுக்கும் இடையிலான வேறுபாடுக்களை அவர்களுக்கு சொல்லிக்கொடுப்பது மிகவும் அவசியமாகிறது. இந்த வகையில் பிரஞ்சு, ஸ்பானிஷ் போன்ற மற்ற மொழிகள் தமிழ் போல் இல்லாமல் கற்றுக்கொள்வதற்குச் சுலபாகவே இருக்கும். மேலும் இந்தவகைப் பாடங்கள் தமிழ்நாடு மற்றும் இலங்கை போன்ற இடங்களில் தமிழ் பயிலும் மாணவர்களுக்குச் சொல்லித்தர வேண்டிய அவசியம் ஏதும் இல்லை.

## 4. மின்வலையில் தமிழ்ப் பாடங்கள்

மின்வலையின் வசதிகள் பெருகிவிட்ட இந்தக் காலக் கட்டத்தில் இது மாதிரியான கடினமான பாடங்களையும் சிறந்த முறையில் கற்றுக் கொடுப்பது மிகவும் எளிதான ஒன்றாக இருக்கிறது. ஒலி, படங்கள், ஒலிப்படங்கள், தேர்வு கொடுப்பதற்கான வழிவகைகள் முதலான கொண்ட தமிழ்ப்பாடங்களை மின்வலை மூலம் அமைத்துத் தருவது மிகவும் எளிதான ஒன்றாக இருக்கிறது. பென்சில்வேனியாப் பல்கலைக் கழகத்தின் தமிழ் வலைப் பகுதி மூலம் தமிழ் எழுத்துக்களை எப்படி எழுதுவது, தமிழ் சொற்களை எப்படி அமைப்பது போன்ற இளநிலை தமிழ்ப் பாடங்கள் முதல் எவ்வாறு தமிழ் பேசுவது, தமிழில் உள்ள வழக்குத் தொடர்கள் முதலான முதுநிலைத் தமிழ்ப் பாடங்கள் வரை பல பாடங்கள் கொடுக்கப்பட்டிருக்கின்றன. இத்தோடு பேராசிரியர் ஹெரால்டு ஷிப்மேன் அவர்களது பேச்சுத் தமிழ் இலக்கணம் முழுவதையும் சுலபமாகப் படிக்கும் வகையில் மின்வலையில் கொடுத்திருக்கிறோம். இவ்விலக்கண நூல் எழுத்துத் தமிழிலிருந்து பேச்சுத் தமிழுக்கு வாக்கியங்களையும் சொற்களையும் மாற்றுவதற்கான விதிகளை வரிசைப் படுத்துகிறது. மேலும் Conversational exchanges என்ற பக்கத்தில் பேச்சுத் தமிழ் வாக்கியங்களையும் அவற்றுக்கு ஈடான எழுத்துத் தமிழ் வாக்கியங்களையும் உடனுக்குடன் மாற்றுவதற்கும் அவற்றை ஒலிவாங்கி மூலம் கேட்பதற்கும் வசதி செய்யப்பட்டுள்ளது. இவ்வலையில் உள்ள பாடங்கள் பின்வருமாறு வடிவமைக்கப்பட்டிருக்கின்றன.

**இள நிலைத் தமிழ்ப் பாடங்கள்**

(Elementary Materials)

தமிழ் எழுத்துக்களை எழுத மற்றும் கேட்க வசதி

சொற்களை அமைக்கப் பயிற்சி

உறவு முறைச் சொற்களை மனதில் இருத்திக்கொள்ளும் வகையில் அமைக்கப்பட்ட பாடங்கள்

எளிய வாக்கியங்களை அமைக்கப் பாடங்கள்

**இடைநிலைத் தமிழ்ப் பாடங்கள்**

(Intermediate Materials)

படிக்கும் திறனை வளர்த்துக்கொள்ளும் வகையில் பாடங்கள் மற்றும் அதற்கான தேர்வு வினாக்கள்

(Reading Literary Tamil)

இலக்கணப் பாடங்கள்

(Grammatical Modules)

தமிழ் வினைவகைகள் மற்றும் அவற்றின் வெவ்வேறு அமைப்பு முறைகள்

தமிழ்ச் சொற்களை மனதில் இருத்திக்கொள்ளும் வகையில் அமைக்கப் பட்ட பாடங்கள்

(Vocabulary Acquisition and Review.)

எழுத்துத் தமிழும் பேச்சுத் தமிழும்

(Writing Tamil: differences between Spoken and Written)

**முதுநிலைத் தமிழ்**

(Advanced Spoken and Written Tamil)

திருச்சி வானொலியில் ஒலிபரப்பப் பட்ட சில நாடகங்கள் (படிக்க, கேட்க, மற்றும் மொழிபெயர்ப்பைத் தெரிந்து கொள்ளும் வகையில் அமைக்கப் பட்ட பாடங்கள்).

தமிழ் நாவல்கள் (படங்கள், ஒலி முதலியன கொண்ட பாடங்கள்)

சீர்காழி கோவிந்தராஜன் பாடல்கள் (பாடல்களை படித்துக்கொண்டே கேட்க வசதியுடன்)

இந்தப் பாடங்கள் மாணவர்கள் தங்களது பேசும் திறனையும் எழுதும் திறனையும் வளர்த்துக்கொள்ளும் வகையில் அமைக்கப் பட்டுள்ளன. இருப்பினும் மின்வலை மூலம் மொழியின் எல்லா அறிவையும் புகட்டிவிடமுடியும் என்று கூறமுடியாது. குறிப்பாக மாணவர்களது பேச்சு மற்றும் வாக்கியங்களைச் சரிப்படுத்தப் போதுமான வாய்ப்புக்கள்

மின்வலையில் இல்லை. இந்த வகையில் மின்வலைப் பாடங்களை ஒரு ஆசிரியரின் துணை கொண்டே முழுமையான அளவில் பயன்படுத்த முடியும்.

## 5. முடிவுரை

பென்சில்வேனியாப் பல்கலைக் கழக தமிழ் வலையில் புதிய பாடங்கள் கொடுக்கவும் ஒலிகளுடன் கூடிய தமிழ் - ஆங்கில அகராதியைக் கொடுக்கவும் வேலைகள் நடந்து வருகின்றன. மொழி கற்க மற்றும் சொல்லிக்கொடுப்பதற்கான கூட்டமைப்பு (Consortium for Language Learning and Teaching) மூலமும் மிகக் குறைவான அளவில் கற்பிக்கப் படும் மொழிகளுக்கான சங்கம் (Less Commonly Taught Languages) மூலமும் பென்சில்வேனியாப் பல்கலைக் கழக தமிழ் மின்வலையை செழுமைப் படுத்துவதற்காக நிதியுதவி கிடைத்துள்ள நிலையில் நாங்கள் வெவ்வேறு நாடுகளில் இருக்கும் தமிழாசிரியர்களிடமிருந்தும் தமிழறிஞர்களிடமிருந்தும் தமிழ்ப் பாடங்கள், சிறு கதைகள், நாவல்கள், ஒலி ஒளிப்படங்கள் (movies) போன்றவை குறித்தான உதவியை நாடுகிறோம். எங்களது தமிழ் வலைப் பக்கத்தை வளப்படுத்துவதில் ஆசிரியர்களின் உரிமைகள் குறித்தான பிரச்சினைகள் பெரும் தடையாக இருக்கிறது. குறிப்பாக இவற்றை எங்களது மின்வலையில் கொடுக்க ஆசிரியர்களிடமிருந்தும் தயாரிப்பாளர்களிடமிருந்தும் பூரண அனுமதி தேவைப் படுகிறது. மேலும் கணிப்பொறியில் தட்டச்சு செய்த பாடங்களோ நாவல்களோ தங்களிடம் இருப்பின் அதை எங்கள் மின்வலையில் சேர்க்கக் கொடுக்க முடியுமெனின் மிகவும் சிறப்புடைத்தாகும். இது குறித்து தாங்கள் எங்களை plc@ccat.sas.upenn.edu என்ற மின் அஞ்சலில் தொடர்பு கொள்ளலாம்.

**கட்டுரை ஆசிரியர் விவரம்:**

முனைவர். அரங்க வாசு R. Vasu

துணை இயக்குனர் (Assistant Director for Multimedia)

பென் மொழிகள் மையம் (Penn Language Center)

714 Williams Hall

பென்சில்வேனியாப் பல்கலைக் கழகம் (University of Pennsylvania)

பிலடெல்பியா, PA 19104 USA

**மின் அஞ்சல்:** vasur@sas.upenn.edu

**மின் வலை இல்ல முகவரி:** http://www.sas.upenn.edu/~vasur

# PROJECT MADURAI FOR E-TEXTS OF TAMIL LITERARY WORKS
## (Madurai Thamizh Ilakkiya Minthokupputh Thittam URL: http://www.tamil.net/projectmadurai/)

*Dr. K. Kalyanasundaram*
*Inst. of Physical Chemistry*
*Swiss Federal Inst. of Technology*
*CH-1015 Lausanne, Switzerland*

**Abstract**

*Project Madurai is an open, voluntary initiative devoted to electronic texts (ETexts) of Tamil literary works. This paper provides a broad overview of the working of this project, Etexts released in the first year of its existence and Etexts that are under preparation. Looking into the future, the paper also raises some of the key issues relating to Tamil EText archives that are to be sorted out, issues such as copyrights of very ancient works, standardization of file format and distribution modes, dedicated OCR for Tamil and an international registry for Tamil Etexts.*

## Introduction

Electronic versions of printed texts (abbreviated as ETexts) of literary works are important pedagogic and scholarly resources. Stored in easily accessible archives, they permit preservation and wider distribution of the literary works around the globe through the means of Internet. It is an elegant approach by which millions of Tamils and Tamil lovers worldwide can access Tamil literary classics right from their home and hence maintain close contact with their culture. With the stated goals of the Tamilnadu Govt. to provide Internet Booths all over Tamilnadu, millions of Tamils living in thousands of towns can also access to the Tamil Literature. Etexts of literary works also allow quick search for phrases, words, and combinations of words in any literary work. This in turn allows one to try out notions and hypotheses with great speed and encourage large-scale searches over oeuvres, genres, and centuries, which are difficult and time-consuming with printed texts alone. Also such work can be done remotely from one's home or office. Thus they facilitate enormously the work of researchers involved in the study of evolution of language and literature.

There are many projects currently active that attempts to put in electronic form ancient literary works. The following are some of the major Etext archives covering English language literature : Oxford Text Archives, On-line Book initiative, Internet Wiretap, Project Gutenberg, Alex Catalog of Etexts, Project Bartleby. Major University Libraries in the US (UVirginia, UCBerkeley, UChicago, UMichigan, UNorth Carolina) even have dedicated Etext Centers. Etexts of several thousand works are available (GB of archived files). Coverage is not restricted to English. There are growing collections for French, German, Italian, Dutch, Scandinavian, Japanese, Chinese, Greek, Hebrew, Russian languages, to mention a few. Except possibly for Sanskrit (for which there are major collections at Utah, Princeton U and Cambridge U), Etext projects for Indic Languages are only symbolic. They are mostly voluntary efforts of few individuals.

## Tamil Etexts and Etext Archives

One can cite several reasons as to why Tamil Language badly needs an extensive public access Etext Archive. As an ancient language dating back to at least two thousand years, it has a very rich repertoire of literary contributions. Unfortunately, a good part of these works remained in palm-leaf manuscript level and never unpublished in printed form. 20th century Tamil literature is largely dominated by novels and associated decrease in the sale of printed copies of ancient literary works covering other domains. With dwindling market, key publishing houses such as Saiva Siddhantha Trust are dropping plans of reprinting classics. A good number of old works have not been reprinted for several decades and hence out of print.

Comprehensive collections of Tamil works ever published are now restricted to a handful of libraries both within and outside India. Due to nature of the printing paper used in olden days and poor storage conditions, many of the published books are being eaten away by insects. The burning out of the entire book collections of major Tamil library of Jaffna in a fire accident is a stark realization of major risks that Tamil faces in loosing its rich literary treasures confined to few places. The following scrolling logo that we use in the Project Madurai summarizes succinctly the benefits of an Etext archive: "kAlatAl aziyAtatu, koTutAl kuraiyAtatu, centaNalAl vEkAtatu mintokuppu. tamizh ilakkiyangkalai nanrE kAttu, ulakengum parappa inrE cEruviir maturai tamiz ilakkiya mintokuppu tittam". We all have a moral obligation to preserve our rich heritage and ensure that the future generation do have access to literary works, possibly via better means of archival and world-wide distribution.

Dr. Thomas Malten of the Univ. of Cologne in Germany possibly is the pioneer to engage in preparation of Etexts of Tamil literary works. Over a decade, he and his colleagues at their Institute of Indology and Tamil Studies prepared Etexts of practically all major works of the Sangam period. Before the advent of modern Tamil computing directly in Tamil scripts, transliterated format was the mode of writing on computers and the Etext archives of Cologne are also in this transliterated version. The Etexts themselves have not made available to the general public (till today). However WAIS-based word search of the archived texts have been available through gopher and Web servers for many years. In collaboration with Prof. George Hart of UC Berkeley, Dr. Malten is currently working on a major Project called "PONGAL 2000" through which their Etexts collections will be made available (hopefully in Tamil Script). The preparation of Etext of Naalayira divya prabhandam (in transliterated format) in early nineties by a group of volunteers led by Prof. P. Dileepan can also be cited as one of the early successful examples of Internet-based voluntary efforts in this domain.

With the availability of free Tamil fonts in the Internet starting from mid-nineties, there has been a phenomenal increase in the amount of Tamil-related materials in the Internet. Through my Tamil Electronic Library web site launched in early 1995, I have been distribute Mylai font free for use on Windows, Mac and Unix platforms. Along with the font, I distribute Etext files of small literary works, particularly of devotional type such as thirukural, kandar shasti kavacham, auvaiyar works, select collections of thevaram, etc. Through postings in the soc.culture.tamil (USENET) and tamil.net (Email discussion list), many expressed the desire to start a collective effort, targeting ancient literary classics. Based on recent calls in tamil.net, this initiative officially took off on the Pongal day of Jan. 14, 1998! After some collective discussions, "Project Madurai" (PM) was taken up as the name to describe this Internet-based Voluntary Initiative. (The abbreviation PM is used in rest of this paper to refer to Project Madurai.). Madurai has been a citadel of Tamil culture. It was Pandyan Kings during their long reigning period, who set up Sangams (academies) for the encouragement and criticism of Tamil Studies. Hence it is appropriate that an Initiative devoted to preserving and distributing Tamil literary works in electronic form is named after this historic city.

## Modus Operandi of Project Madurai

Project Madurai is based on voluntary cooperation between lovers living in several countries. It is planned to use both direct text input and OCR of scanned images to generate Etexts. In the absence of a dedicated OCR, currently Etext preparations are through direct input by the PM volunteers. (Univ of Chicago recently opted to microfilm the entire 50000+book collections of Raja Muthiah Library of Chennai. Using a dedicated OCR system, it should be possible to generate the Etexts from such microfilms in a short span of time!) The Etexts are subsequently proof-read by some other PM volunteer before they are archived and released to the general public. Completed Etexts are made available through a dedicated web site <www.tamil.net/projectmadurai/ > hosted by Asia Pacific Internet Company (APIC) of Sydney, Australia. The Etexts are distributed in plain text format and also as formatted texts in the form of Web pages (HTML version). Portable Document Format (PDF) Distribution is becoming increasingly popular as a mode of distribution of formatted texts in the Internet. Free Acrobat readers are available to read pdf files in all commonly used computer platforms. Distribution of Tamil Etexts in pdf format has just begun for select works. When the Etext collections become substantial, it is planned to make them available in CD-ROMs.

As a grass-root Internet-based effort devoted to free distribution of Tamil Etexts, PM belongs to any one committed to the stated goals of the Project. Entire coordination and exchange of files are through standard Email. It suffices to have a simple Email connection for anyone to be engaged in PM activities and there is no need to have Internet connection. Operating rule are set by the volunteers themselves. Project execution and evolution are discussed routinely through a dedicated Email discussion list <pmadurai@onelist.com>. The project is neither commercial nor governmental.

Considerable flexibility is given to the volunteers in the choice of the work he/she chooses to do text input and also the font encoding format. The goal is to let the volunteers work in an environment (font and computer) he/she feels comfortable with and the work where there is some personal interest. Very minimal constraints if any will be imposed on the volunteers who will do the major task of keying in of texts. Appropriate acknowledge is given to the volunteers involved in the web pages announcing the availability of a given Etext and also in source files. In addition to the preparation of Etexts, PM volunteers are also involved in related activities such as development of a dedicated OCR for Tamil, softwares for inter-conversion to other font encoding formats etc. Through these converters it is possible to provide Etexts in at least five different font encoding formats (typewriter, romanized/transliterated, Adhawin, Inaimathi/Anjal and Mylai). This renders the access of Etexts to a much larger Internet user community.

## Font Encoding for Etexts

In Project Madurai, we are committed to using an internationally recognized font encoding standard for the archives. In the first year, we chose to work with following two font formats for primary file archiving: Inaimathi/Anjal and Mylai. The reasons for limiting the choice to these formats were the following: a) Over 90% of the Tamils worldwide use one of these; b) fonts are available free for use on all of the three major computer platforms - windows, Macintosh and UNIX. Anyone who uses these computers can work in all of these formats; and c) converters such as Adhawin and Anjal are available that work reliably to go between these formats and to several others mentioned above. So we can provide equivalent form of Etexts in typewriter and romanized/transliterated formats to interested parties. An important added component to the Etext is an introductory note (presented as a Web page) that sets the background for the work and author. It is written in a simple style so as to be easily understood by native Tamils and also by language lovers/researchers.

## Selection of Works for Etext Archives

Project Madurai is deeply committed to respect Copyright protection given to authors of literary works. Even though the copyright rules vary from country to country, in most of the Etext archiving projects, an elapse of at least 75 years after the life span of the author is considered a safe criterion before the work can be considered to be in public domain. (In the USA, Sonny Bonno Bill approved by the Congress late in 1998 has extended the author protection to a total of 95 years). So, as a rule of thump, we can consider works of authors of 19th century and earlier dates to be in public domain. Selection of works for Etext preparation is primarily dictated by this date constraint. Hence is the tilt for archiving ancient literary classics. Another reason for going for ancient literature is that most of them are out of print and stand the risk of getting lost to the world.

As regards the literary works of the 20th Century, efforts are being made to contact the authors (or their legal heirs) or publishers to obtain written consent for the work to be distributed in electronic form free on the Internet. Through special arrangements with the nearest kins, Tamilnadu Govt. has placed "in public domain" the works of select recent authors - Bharathiyar, Bharathidaasan, Annadurai and others. Preference is given to coverage of works of these authors in the first phase of the project. Some Etext archives do accept shareware works where the user has to pay a fixed amount to download a copy. Since this will involve administrative bureaucracy, we will not go for shareware option, at least in the initial stages.

With the above broad guidelines, target works for Etext preparation are chosen. Present choices can be described by the following four categories:

I. Works of popular appeal and devotional literature:

thirukuRaL of thiruvaLLuvar, works of auvaiyAr (AthicUdi, kondRai vEnthan, nalvazhi, mUthurai, thiruppuhaz of aruNagiri nAthar, thiruvarutpa of rAmalinga adigaL, kandar alangAram, kandar anubhoothi, kandar shasti kavacham, kanthar kavacham, vEl virutham,., songs of cithars such as pattinathAr, songs of thAyumAnavar, ciRappurANam of umarup pulavar, ...

II. Great anthologies of:

works of sangam period (ettuththohai, paththuppAttu, pathinenkiizkaNanku), twelve thirumarais, divya prabhandam, saiva siddhantha sastras, grammar works and their commentaries, kamba rAmAyaNam, and Epics such as cilappathikAram, maNimEkalai,..... Univ. of Cologne Archives has many of these works in transliterated format. Current priority here is to go for works that are not already covered there.

III. Works of 20th Century Authors:

A. Authors whose literary works are already in public domain Tamilnadu Govt. has already placed the literary works of a select list of authors in public domain. An attempt will be made to put in electronic form the works of these authors: C. Subramania BharathiyAr, BharathidAsan, maRaimalai adigaL, Pattukottai Kalyanasundaram and C.N. Annadurai. Very recently Tamilnadu Govt. has announced plans to place in public domain the works of some more distinguished authors such as Kalki Krishnamurthy, Thiru Vi. Ka, Yogi Suddhanantha Bharathi,.... Through the initiatives of one of our volunteer Mrs. Uma Shankar, we have obtained permission to place the works of Kaviyogi Maharishi Suddhanantha Bharathiyar in the Internet.

B. Recent Authors who give permission: 20th Century Literature is largely dominated by novels. One of our PM volunteers Mr. Gandhi Kannadhasan has been lobbying with leading novelists, encouraging them to place at least one or two representative novels in the electronic archives. The following is a short list of modern authors who have expressed willingness to participate in this process of putting select works in Etext form: Kannadhasan (through legal heirs), Akilan, Naa Paarthasarathy (through legal heirs), Prabanjan, Indumathi, Rajam Krishnan, Indira Parthasarathi, Malan and Jeyakanthan.

IV. Translations and Commentaries of Literary Works:

Understanding of our literary works are assisted by Commentaries and Translations in English and other world languages. There are many classics that are in "public domain". The English translations of Rev. G.U. Pope (1828-1908) of thirukuRaL and thiruvAcakam, parimElazhakar Urai of thirukuRaL are examples. Project Madurai would attempt to provide authentic Etext versions of these ancient but classic works. Arrangements are being made to put the English translation of thirukuRaL and other literary works of Yogi Suddhantha Bharathi.

V. Literary works of Authors of Srilankan Origin:

Jaffna Library has been the major resource center for the literary works of authors of Sri Lankan Origin. Unfortunately fire at this renowned library destroyed all the precious book collections. Majority of authors of sri lankan origin are also displaced from their homeland. There is a major risk of loosing all the works of sri lankan origin. Hence Project Madurai will make an effort to bring in electronic form whatever work we can have access. Thanks to the initiatives of Mr. Padmanabha Iyer of London and Dr. N. Kannan of Germany we have already a couple of anthologies ready for release.

## Worldwide Network of Volunteers

When the project was launched a year ago, we started with a small group of about 25 volunteers interested to get involved. In one year the ring of volunteers has expanded to count 80 and they all come from four corners of the world: U.S.A [from the states of ca, nm, tx, ka, il, in, ga, oh, vi, ny, nj and nh], Canada [states of quebec and regina], Western Europe [London (UK), Cork(I), Kiel (G), Lausanne (CH), Helsinki (Fi)], Gulf States [Riyadh, Abu Dhabi], India [from Chennai, Trichi, Katpadi,

Coimbatore, Bangalore, Bombay, Delhi, Baroda, Kanpur], Sri Lanka [Colombo], Southeast Asia [Singapore, Malaysia] and Australasia [from cities of Sydney, Wellington and Auckland].

Etexts Completed as of Date

The following is a list of works for which we have Etext preparation nearly complete (many of them indicated by * already released to the public) thirukuRaL*, auvaiyar works*, thiruvaachagam*, thirumanthiram*, naalayira divya prabhandam*, thaNNir dEsam* of Vairamuthu, Bharathiyar songs-part I*, dEsika prabhandam* of vEdAntha dEsikar, Bharathiyar Tamil translation and commentary* of Bhagavad Gita, naLaveNbA* of pukazhendip pulavar, nAladiyAr*, 9th thirumurai (thiruisappas), 11th thirumurai covering works of kAraikAl ammaiyar, nakkiirar, kapilar, nambi AndAr nambi, English translation of ThirukuRaL by Yogi Suddhanantha Bharathi*, anthology of contemporary literature of Tamil immigrants (particularly those of Sri Lankan origin), Holy Bible - New Testament in Tamil (part I)*, thiruvarutpa of rAmalinga adigaL and pathiRRupaththu

Etexts in Preparation

The following is a short list of Tamil works that are currently under preparation: paadalkaL of ciddhars pattinathAr, pathirakiri and others, works of Bharathidaasan, thiruppuhaz of aruNagiri, ciRappurANam of umarup pulavar, English translation of thiruvAcakam by Rev GU Pope, Tamil commentary of thirukuRaL by parimElazhakar, nAttuppurap pAdalkaL /Tamil folk songs, kamba rAmAyaNam, Tamil translation of Kalevala (great Finnish Epic) by UthayaNan, Modern Tamil Writings of Sri Lankan Authors in Exile, Bharatha Shakti Maha Kaviyam and other works of Yogi Suddhanantha Bharathi, select works of Kavinjar Kannadhasan, Thampikki aNNAvin kadithangaL of CN Annathurai, cilappathikaaram,...

## General Issues Related to Tamil Etext Archives

As part of this presentation, I would like to discuss some important issues related to Tamil Etext Archives hoping to find some response from the participants of this conference and also from the Tamilnadu Govt. First issue is related to the question of Copyrights: Which of the ancient works are in the public domain? It was mentioned earlier that Berne Convention of Copyrights signed by many countries (including India) confers protection to authors (and their legal heirs) to 75 years after the passing away of the author. In USA Sonny Bonno Bill has extended this limit to 95 years. As per this reference, works of authors who left us by the beginning of this 20th C are in public domain. In Tamil, there have been contradictory reports that even works of Sangam period are still under copyright. Promoters of Etext archives are very much concerned by claims of this kind and have delayed public release of Etexts for several years. Herein I would like to propose some suggestions, particularly to works that date to 19th Century or earlier periods. I must emphasis that there are purely my personal views and I am not a legal expert to make definitive statements.

A significant number of works dating back to several centuries were never published in the printed book form that we are familiar with. Most of them were at the palm-leaf manuscript level. Only in recent times there have been systematic efforts to publish these very ancient works of Tamil. Thamizh thaththA U.Ve. CuvAminAtha Aiyar was a pioneer in this area and has been responsible for several critically edited versions of Sangam period works. Clearly such critically edited, recent publications of ancient works are subject to copyright, even if the original work may date several hundred/thousand years back. Maiden publications of hitherto unpublished ones (for e.g. works that are still at the palm-leaf manuscript stage) are also subject to copyright by the recent editor and publisher. Institute of Asian Studies, for example, is an Institution committed to publishing ancient works from palm-leaves. Recent commentaries of ancient works are also subject to copyright, so also the recent translations to English and other languages of ancient works. On the other hand; very ancient works such as thirukuRaL that have been repeatedly reprinted by several publishers during this century are possibly in public domain. There should also be no constraints for someone to extract the source/moolam verses of ancient works from recent commentaries and public the original work alone in Etext form. The tasks of promoters of Etext Archives will be largely facilitated if the Tamilnadu Govt. appoints a Special Experts Committee to look into this question. Working along with the Book Publishers Association of Tamilnadu, they make definitive statements on which of the ancient works are indeed in public domain for anyone to freely reproduce electronically or otherwise.

A second important issue is Standardization of the file/data format in Etext Archives. There are several isolated efforts currently underway all aimed at archiving works electronically. In the absence of any clear standards, each archive is using its own font-encoding and file storage formats. Institutions such as Univ of Pennsylvania interested in Linguistics are interested in "Corpus" text where every word in the Etext is tagged. In view of the growing popularity of the Web, formatted texts are made available in HTML formats. Librarians, on the other hand prefer that such formatted texts are in SGML format. Efficient working of Internet-based search engines require some standardisation implemented in these archives.

Transliteration scheme is another area where standardization needs to be enforced. Etext archives of works in transliterated format is still of interest to researchers of non-Tamil origin. On the formats for transliteration, the classical scheme using diacritical markers (dots and bars below and above) is used by Library of Congress in its cataloguing and also by the Inst. of Asian Studies in their Lexicon and publications of palm-leaf manuscripts. In view of the practical problems in using such a scheme in electronic communication/exchanges, alternate schemes based on plain ASCII characters (with or without upper case characters) are being used increasingly. At least five different popular schemes in use. Here again the Special Committee of the Tamilnadu Govt. can play important role in setting standards.

A dedicated Optical Character Recognition (OCR) package for will assist enormously the rapid processing of scanned images (even those from microfilms) and generation of Etexts. Currently we do not have any such OCR software for Tamil. Urgent need for such an OCR was recognized in the last TamilNet'97 conference held in Singapore but till date, there has not been any systematic follow up. In PM, we have a group of volunteers (all computer software professionals) are working to develop one such OCR package and make it available free to the general public. With growing interest to Tamil Computing, it is likely that commercial softwares will appear soon in the market. Voluntary efforts such as that of PM can benefit enormously if a polyvalent OCR is made available free.

Cooperation and coordination between all individual and institutional efforts engaged in Etext preparation are necessary to avoid duplication in the preparation of Etexts. A central/international registry of Etexts (public access or otherwise) can help the cause. A central/international registry of new projects of Etexts being undertaken (where anyone can find out if someone is working on a given work) is also helpful in coordination. During its first year, in Project Madurai we have found many committed lovers who are willing to get involved in preparation of Etexts of works. Most of them are based overseas and they do not have access to a printed copy of the work to start the text input process. A Chennai-based Book Bank, for example, can help the PM cause and speed up the process.

## Concluding Remarks

Project Madurai is a concrete example of how like-minded souls can work collectively for a common cause even if they are physically located in four different corners of the world. Email communication is a marvelous gift of the electronic age and we are successfully using it. For many of us, engagement in PM efforts rekindled our interest in Language and its rich Literature (culture in general). Millions of Tamils living outside India do have access to Internet already and many of them are thrilled by the fact that they can go to Internet and download Etext files of Literary Classics. Particularly for those placed in remote cities/countries far off where they cannot easily get printed copies of books even if they want to, Etext archives are wonderful gifts. In this aspect, Etext projects such as PM do fulfill their stated goal of promoting preservation of Culture worldwide. In conclusion, I would like to record my deep appreciation and gratitude to all the PM volunteers for their enthusiasm and hard work.

# COMPUTER - AIDED IDENTIFICATION OF TAMIL WORDS (CAITW) (BASED ON NANNOOL PHONOLOGY)

*Mr. Va. mu. se. Kaviyarasan, M.Sc. M. Tech.*
*Assistant General Maneger (Systems)*
*V.T.I. House, Rajaji Salai, Chennai - 600 001*
*and*
*Dr. Va. mu. Se. Andavar*
*Lecturer in Tamil*
*C. Kandasamy Naidu College for Men*
*Anna nagar, Chennai - 600 102*

## Introduction

There are many research works carried out by Linguistics as well as by Computer Scientists relating language and Computer Science under the discipline called Natural Language Processing (NLP).

The present project deals with the problems faced in identifying Phonologically Possible Tamil Words based on Nannool.

The first part briefly explains some of the basic notions in Natural Language Processing (NLP) necessary to understand the present work.

In the second part, an introduction to the Tamil Grammar "Nannool" with special reference to the Phonological rules, is given.

The structure of the Programme is described in the third part.

The usage and the limitations of the present Programme are summarised in the final part.

### *Natural Language Processing (NLP)*

The research relating Language and Computer can be classified into two groups.

1. Using Computers for Language research.
2. Natural Language processing for Articificial Intelligence(AI).

### *Using Computers for Language Research*

In Language research Computers can be used by Linguistis for various Processing and analysis, as used by the researchers in the other academic fields.

This research may enable the Linguists in carrying out their linguistic research more accurately and very fast.

### *Natural Language Processing*

At present, Computer Scientists use artificial languages such as BASIC, C etc., However, they attempt to use the human Natural Language. This achievement in the Computer field is called NLP. This may promote the Artificial Intelligence (AI) research in Computer Science.

In this NLP research, Computer Scientists and Lingustics are equally interested. They have been undertaking this research as an interdisciplinary one.

The main problem to be solved in NLP is to accommdate grammer formalisms into Computer formalisms. This kind of attempts are being made by several Computer Scientists. The following are such formalisms – TEAM ELIZA etc,.

NLP can be done in the following stages.

1. Phonology
2. Morphology
3. Syntax
4. Semantics and Pragmatics.

In this research, the problems related to the conditioning of Phonologically possible Tamil Words, based on Nannool, is being discussed.

***Nannool***

The earliest Tamil Grammer available is Tolkappiam written by Tolkappiar dating back to 5th century B.C. It contains about 1611 rules. Nannool is one of the well known Grammar after Tolkappiam. The author of Nannool is a Saint called Bhavananthi. It belongs to13th Century A.D. There are about 462 rules in Nannool. Each rule is expressed in a varse like form called "Sutra". Even though Nannool is said to contain the following five chapters. 1. Elutu 2. Col 3. Porul 4. Yappu 5. Ani. At present only two chapters only Elutu and Col are available.

***Eluthu Athikaram Contains the following five chapters.***

1. எழுத்தியல் – *(eluttiyal)*
2. பதவியல் – *(pataviyal)*
3. உயிர் ஈற்றுப் புணரியல் – *(uyir irrup punariyal)*
4. மெய் ஈற்றுப் புணரியல் – *(mey irrup punariyal)*
5. உருபுப்புணரியல் – *(urupu punariyal)*

***Col athikaram contains the following five chapters.***

1. பெயரியல் – *(peyariyal)*
2. வினையியல் – *(vinaiyiyal)*
3. பொதுவியல் – *(potuviyal)*
4. இடையியல் – *(itaiyiyal)*
5. உரியியல் – *(uriyiyal)*

Elutiyal deals with the Various Tamil Speech Sounds and graphemes existing in Tamil. Rule no 74 to 97 of Nannool describes the Speech Sounds based on position and manner of Articulation.

Rule No. 98 to 101 deals with the length of the Sounds ("mattirai").

Rule No. 102 to 121 of Nannool describes the first, final and intermediate phonemes of Tamil Words. By transforming these grammatical formalisms in Nannool in to Computer formalisms in a Programme designed in the **Programming Language 'C'** it is attempted here to identify the Phonologically possible Tamil Words.

The details of this programme is discussed in the following.

The following Nannool rules are used in the Programme.

***Mutal Xialy.***

Rule: 102. Pannl ruyruf kacataxa pamavaya. Fetavl rlxtuyir meyyum mozimutal.

Rule: 102. uUoO alaVOdu ammudal.

Rule: 104. aA uU Ov yammudal.

Rule: 105. aA eo VVo dAkulFammudal.

Rule: 106. Cuddiya ekara VinAvazi aVVl.

***IRRU Xilay.***

Rule: 107. AVi FNaXama nayarala vazala mey
CAyu mukara XALARu MIRE.

***"Kokkudil" spell conditioner***

a. Usually, the programme called "Spell Checker" contains two parts

1. Identification of Errors.

2. Correction of Errors.

In the Programme "Kokkudil" Spell Conditioner, only the indentification of errors is being considered. This "Kokkudil" Spell Conditioner conditions the Phonologiocally possible Tamil words based on the Phonological rules expressed in "Nannool"

b. The Tamil phonemes are represented by the following trasnlitration System in the programme as given below.

**Vowel - Chart.**

| | | | | | |
|---|---|---|---|---|---|
| அ - a | ஆ - A | இ - i | ஈ - I | உ - u | ஊ - U |
| எ - e | ஏ - E | ஐ - Y | ஒ - o | ஓ - O | ஔ - V |

**Consonant - chart**

| | | | | | |
|---|---|---|---|---|---|
| க் - k | ச் - c | ட் - d | த் - t | ப் - P | ற் - R |
| ங் - f | ஞ் - F | ண் - N | ந் - X | ன் - n | ய் - y |
| ர் - r | ல் - l | வ் - V | ழ் - Z | ள் - L | ம் - m |

The first part of the programme indentifies the Phonemes uEIR (Vowel), Mei (Consonant), and uEIRMei (syllable or consonant+Vowel) according to Tamil tradition.

| | | |
|---|---|---|
| for example | : | Kavi |
| Tamil Script | : | Ka |
| Tamil Script | : | Vi |

"Ka" is a valid first Phoneme. Nannool Rule No. 102, 103.

"i" is a valid Last Phoneme. Nannool Rule No. 107.

An the Second part indentifies the errors.

for example : ravi

Tamil Phoneme : ra

Tamil Phoneme : vi

"ra" is not Valid first phoneme. Nannool Rule No. 102.

***Summary***

This "Kokkudil" Spell Checker can be considered as the Programme of Phonological rules in Nannool as well as the programme Conditioning Phonologically Possible Tamil Words based on Nannool.

This Programme can easily be further extended with some modern variations to make it a complete one.

By adding the lexical print necessary for correction of these errors, this "Kokkudil" Spell Conditioner can be made as a Complete Tamil Spell Checker.

The crucial point to be noted is that the Nannool Grammer formalisms are very close to the Computer formalisms and if makes the computerisation easy.

# கணிமொழி - 'சி'

(B.E.மாணவத்திட்டப்பணியின் விளைவு)

Dr.மு.பொன்னவைக்கோ
தலைமைப் பேராசிரியர், கணிப்பொறித்துறை
கிரசன்ட் பொறியியல் கல்லூரி
சென்னை - 600048.

இந்நூற்றாண்டின் ஈடு இணையற்ற கண்டுபிடிப்புகளில் கணிப்பொறி முதல்நிலை வகிக்கிறது. ஆங்கில மொழி அனைத்துப் பயன்பாட்டிலும் மிகுந்த ஆதிக்கம் செலுத்திவருகின்றது. இதனால் ஆங்கிலம் அறியாத நூற்றுக்கணக்கான மக்கள் கணிப்பொறியை இயக்கும் வாய்ப்பினை இழந்து வருகிறார்கள். தமிழ் மொழி அறிந்தவர்களும் எளிய முறையில் கணிப்பொறியை இயக்கிப் பயன்பெற தமிழ் மென்பொருள் தெவைப்படுகின்றது. அசுர வேகத்தில் வளர்ந்து கொண்டுள்ள இந்த நேரத்தில், இதுவரை ஆங்கில வரிவடிவங்களைக் கொண்டு வடிவமைக்கப்பட்டுள்ள அத்துணை மென்பொருட்களையும் தமிழில் படைப்பது எளிதான செயலன்று. ஆயின் ஆங்கில மென் பொருட்களுக்கு முன்னிலை செயலாக்கிகளை தமிழில் வடிவமைப்பது எளிது. இந்த அடிப்படையில் 'C' என்ற கணிப்பொறி மொழிக்கு முன்னிலை செயலாக்கியை (Preprocesser) வடிவமைக்கும் ஆய்வின் விளைவே இக்கட்டுரையாகும். இக்கட்டுரையில் 'C' கணிமொழிக்கான முன்னிலை செயலாக்கியின் வடிவமைப்புகள் விவரிக்கப்பட்டுள்ளன.

## உருக்கணம் (Character Set)

ஒரு மொழியின் முக்கிய அங்கங்களாவன : (1) உருக்கள் (Characters), (2) தொடரியல் (Syntax), (3) சொற்பொருளியல் (Semantics).

கணிமொழி - சி - யின்

உருக்கள் கீழ்கண்ட வகையில் வகையறுக்கப்படுகின்றன:

1) எழுத்துக்கள்
2) இலக்கங்கள்
3) சிறப்புக் குறியீடுகள்

**எழுத்துக்கள்**

அ முதல் ஃ வரை
க் முதல் ன் வரை

**இலக்கங்கள்**

0 முதல் 9 வரை
௧ முதல் ௯ வரை

**சிறப்பு குறியீடுகள்**

' " $ [ . ! # ] ; | & {
: \ ^ } ? ~ ( , - )

**தரவின் வகைகள்**

1) முழு எண்
2) இயல் எண்
3) உரு

**தரவு வகையின் அறிவிப்பு கூற்றுகள்**

தரவு வகை அறிவிப்பின் மூலம் மாறியின் பெயர் மற்றும் அதில் தேக்கிவைக்கப்பட்டுள்ள தரவின் வகை, தொகுப்பிற்கு கூறப்படுகிறது. நிரலில் (Program) உபயோகப்படுத்தப்படும் மாறியானது அதன் பயன்பாட்டிற்கு முன் அறிவிக்கப்பட வேண்டும்.

**தொடரியல் (Syntax)**

தரவு வகை க1,க2, .......... கா;

**தரவு வகை அறிவிப்பின் எடுத்துக்காட்டுகள்**

முழு எண் க
இயல் எண் வ
உரு ச

**இயக்கிகள் (Operators)**

இயக்கி எனும் குறியீடு கணித (Arithmetic) மற்றும் ஏரண(Logic) நிகழ்வுகளைக் கையாளப் பயன்படுகிறது. இயக்கி கணித, ஏரணக் கோவைகளில் முக்கிய பங்கு வகிக்கிறது.

**இயக்கியின் பல்வேறு கூறுகளாவன**

1) எண்கணித இயக்கி
2) உறவுநிலை இயக்கி
3) ஏரண இயக்கி
4) மதிப்பளி இயக்கி

**எண்கணித இயக்கிகள்**

+, -, *, /, %

**உறுவு இயக்கிகள்**

<, >, <=, >=, ==, !=

**ஏரண இயக்கிகள்**

&&, ||, !

**மதிப்பளி இயக்கி**

=

**கோவையின் மதிப்பீடு**

மதிப்பளி கூற்றின் வாயிலாக கோவையின் மதிப்பீடானது அறியக்கூடும்.

மாறி = கோவை

**எடுத்துக்காட்டு**

க = ம * வ - த

**உள்ளீடு / வெளியீடு கூற்றுகள்**

கணிப்பொறி நிரலின் மூன்று முக்கிய செயல் கூறுகளாவன

1) படித்தல்
2) செயல் படுத்தல்
3) எழுதுதல்

உள்ளீடு / வெளியீடு இயக்கத்தில் உள்ளீடானது நிலையான உள்ளீட்டுத் தொகுப்பிலிருந்து உருவினைப் படிக்கிறது மற்றும் வெளியீடானது நிலையான வெளியீட்டுத் தொகப்பில் உருவினை எழுதுகிறது.

**உள்ளீடு கூற்றின் தொடரியல்**

படி ("%மு" , &க);

இக்கூற்றின் மூவம் ஒரு முழுஎண் நிவையான உள்ளீட்டுத் தொகுப்பிலிருந்து படிக்கப்படுகிறது.

படி ("%இ" , &க);

இக்கூற்றின் மூலம் ஒரு இயல் எண் நிலையான உள்ளீட்டுத் தொகுப்பிலிருந்து படிக்கப்படுகிறது.

படி ("%உ" , &க);

இக்கூற்றின் மூலம் ஒரு உரு நிலையான உள்ளீட்டுத் தொகுப்பிலிருந்து படிக்கப்படுகிறது.

**வெளியீடு கூற்றின் தொடரியல்**

அச்சிடு("%மு" , &க);

இக்கூற்றின் மூலம் ஒரு முழுஎண் நிலையான வெளியீட்டுத் தொகுப்பில் எழுதப்படுகிறது.

அச்சிடு("%இ" , &க);

ஒரு இயல் எண் நிலையான வெளியீட்டுத் தொகுப்பில் எழுதப்படுகிறது.

அச்சிடு("%உ" , &க);

இக்கூற்றின் மூலம் ஒரு உரு நிலையான வெளியீட்டுத் தொகுப்பில் எழுதப்படுகிறது.

## கட்டுப்பாட்டுக் கூற்றுகள்

நிரலிலுள்ள கூற்றுகள் வரிசை முறையில் செயல்படுத்தப்படுகிறது. சில நேரங்களில் கூற்றுகளின் வரிசைமுறையை மாற்றி செயல்படுத்துவதற்கு கட்டுப்பாட்டுக் கூற்று பயன்படுகிறது.

கட்டுப்பாட்டுக் கூற்றின் இனங்களாவன

1) எனில் கூற்று (IF Statement)
2) எனில் ......... இல்லெனில் கூற்று (If Then .... Else Statement)
3) செல் கூற்று (Go To Statement)
4) நிலைமாற்றுக் கூற்று (CASE Statement)

### 1) எனில் கூற்று

எனில் கூற்று இரண்டு வகையான முடிவுகளை எடுக்க உதவுகின்றது. ஒரு நிபந்தனையை சரிபார்த்து ஆம்/இல்லை என்பதின் அடிப்படையில் ஏதாவதொரு கூற்றினை அடையும்.

எனில் கூற்றின் தொடரியல்

```
எனில் (நிபந்தனை)
{
        __________
        __________

}
```

### 2) எனில் ...... இல்லெனில் கூற்று

ஒரு நிபந்தனையை சரிபார்த்து, ஆமெனில், எனில் கூற்றை தொடரும் கூற்றுகளை செயல்படுத்தும். இல்லையெனில், இல்லெனில் கூற்றைத் தொடரும் கூற்றுகளை செயல்படுத்தும்.

எனில் ....... இல்லெனில் கூற்றின் தொடரியல்

```
எனில் (நிபந்தனை)
{
        __________
        __________

}
இல்லெனில்
{
        __________
        __________

}
```

### 3) செல் கூற்று

ஒரு நிரலில் நிபந்தனையற்ற இடமாற்றத்திற்கு செல்கூற்று உதவுகிறது. ஓர் இடத்தினை அடைவதற்கான அடையாளம் செல் கூற்றின் வாயிலாக வழங்கப்படுகிறது.

செல் கூற்றின் தொடரியல்

செல் அடையாளம்;

____________

____________

அடையாளம்: கூற்று;

4) நிலைமாற்றுக் கூற்று

எனில் கூற்றில் பல பிரிவுகள் வருமாயின் நிரலின் சிக்கல் மிகுதியாகிறது. ஆகையால் நிலைமாற்றுக் கூற்று உருவாக்கப்பட்டது. நிலைமாற்றுக் கூற்று என்பது பல்வழி முடிவு காண் கூற்றாகும்.

ஒரு மாறியின் மதிப்பையோ அல்லது கோவையின் மதிப்பையோ, மதிப்புத் தொகுதியோடு ஒப்பிட்டு, ஏதேனும் ஒரு மதிப்பு பொருந்துமேயானால் அந்த கூற்றுத் தொகுப்பை செயல்படுத்தும்.

நிலைமாற்றுக் கூற்றின் தொடரியல்

```
நிலைமாற்று (கோவை)
{
        நிகழ் 1 :
               ________
               ________
               முறி ;
        நிகழ் 2 :
               ________
               ________
               முறி ;
        தவிர :
               ________
               ________
}
```

பன்முறை செய் கூற்றுகள்

பன்முறை செய் கூற்றில் ஒரு நிபந்தனை முடிவுறும் வரை உள்ள கூற்றுகள் வரிசை முறையில் செயல்படுத்தப்படுகின்றன. கட்டுப்பாட்டுக் கூற்று சில நிபந்தனைகளை சரிபார்த்து, கூற்றுத் தொகுப்பை பன்முறை செயல்படச் செய்கிறது.

பன்முறைசெய் கூற்றின் இனங்களாவன

1) உண்மையெனில் (While (Condition) ........) கூற்று.
2) செய் ........ உண்மையெனில் (Do ........ While (Condition)) கூற்று
3) ஆக (For ........ ) கூற்று

1) உண்மையெனில் கூற்று

உண்மையெனில் கூற்று நிபந்தனை தவறாகும் வரை கூற்றுத் தொகுப்பினைபன்முறை செயல்படுத்தும்.

உண்மையெனில் கூற்றின் தொடரியல்

உண்மையெனில் (நிபந்தனை)

{

________

________

}

2) செய் ....... உண்மையெனில் கூற்று

செய் உண்மையெனில் கூற்று, கூற்றுத் தொகுப்பினை ஒரு முறையேனும் செயல்படுத்தும்.

செய் ....... உண்மையெனில் கூற்றின் தொடரியல்

செய்
{

________

________

}
உண்மையெனில் (நிபந்தனை)

3) ஆக கூற்று

ஆக கூற்றில், கூற்றுத் தொகுப்பு குறிப்பிட்ட முறைச் செயல்படுத்தப்படுகிறது.

ஆக கூற்றின் தொடரியல்

ஆக (தொடக்க மதிப்பளித்தல்; நிபந்தனை; ஏறுமானம்)
{

________

________

}

மேலே கூறப்பட்டுள்ள கூற்றுகளின் வாயிலாக ஒரு தமிழ் நிரலை எழுதி, தொகுத்து, செயல்படுத்த இயலும்.

நன்றி

'C' கணிமொழிக்கான இந்த முன்னிலை செயலாக்கியை வடிவமைத்து உருவாக்கிக் கொண்டுள்ள மாணவிகள் ஃபிர்தோசு, நசரின் ஃபாத்திமா, சாந்தி, சமீமா, ஃபரிதா பானு ஆகியோர்க்கு மிகுந்த பாராட்டுகளும், நன்றியும் உரித்தாக்கப்படுகின்றன.

***

# SENTAMIL - LEARNING / TEACHING TAMIL THROUGH MULTIMEDIA

*N. Anbarasan, CEO, Applesoft*
*No 39, Ist Cross, Ist Main*
*Shivanagar, W. 0. C. Road*
*Bangalore - 560 010.*
*Tel : 3386167, Telefax : 3357167*
*e-mail: aplesoft@vsnl.com, anbu.arasan@axcess.net.in*

The purpose of presenting this project is to emphasise and demonstrate that the learning materials/ teaching components could be prepared through an innovative multimedia.

## Prologue

### *Education*

Education is the foundation underlying the achievements of mankind. The subject matter of this paper is encouraging education in a stimulated teaching and learning by the higher potent teachers to the less potent taughts. The present Seminar focusses more on this aspect and this paper focusses on the supplementary materials in reducing the dropouts and help the individuals understand the subject with more clarity.

Education is not a one time process but a continuous life long process of man making and society building. Education begins at the womb and end at the tomb but yet leaving behind the achievements as legendaries for making better future for the future generation. A strong foundation can be laid with the help of modern technology, which will help in the life long learning curve.

### *Educational psychology*

During the different stages of growth and development the children undergo various psychological barriers which is clearly addressed by the Educational psychology. The urges and demands need to be taken into account for effective educational means. A smooth and unhindred flow of learning material will help the teacher to supplement his classroom activities in terms of cohesion and relation. Due to the alarming expansion of knowledge horizon the teaching materials should be innovative and improvised. The multimedia based materials are effective and productive oriented as they are based on graphics, animation and interactivity. A homogeneous classroom is not guaranteed as pupils coming from different corners having diversity in language, culture, socio economic, tradition, custom, caste, religion etc. With this background, pupils respond and react differently to the same situation. Multimedia helps in preparing an unumbiguous material. As different individuals require different types of teaching and learning suitable to their individualsm, multimedia has the scope for configuring to meet these requirements. This would ensure better results and expected outcome. Multimedia helps one to get conformed with the pace of progress one maintains and to accelerate the rate of achievement and ensure proper attainment.

### *Tamil*

The only classical language in usage today, is Tamil. Tamil is an ancient Dravidian language used in various parts of the globe. Tamil is a Diglossic language with a wide cleavage between the written and spoken language. The spoken Tamil has many dialectal variations. Since, there is a considerable difference in the spoken and written Tamil, at various linguistic levels / take Tamil as an example and try to highlight the need.for supplementary materials in bridging the gap between home dialect (spoken Tamil) and School language (written Tamil).

### *Tamil learning*

Language learning is a complex and time consuming process. When it comes to Tamil learning pupils come with different socio-economic, caste, religious back grounds. Pupil experience the difference

between Home dialect (spoken language -Tamil) and the school language (written language - Tamil) at various linguistic levels such as phonological, grammatical, lexical and semantic. The dialect is acquired out and out through the socialisation process, where as school language is learned in a formal schooling situation. The theme behind this paper is to make the school language learning interesting and motivate the children and enthuse an aptitude, towards that end.

### *Stages in language learning and the contribution of Multimedia*

The various stages of language development in children are as follow:

**Comprehension** : The power of comprehension develops even before the child begins to use words. The child can comprehend more words than they can speak. The ability of comprehension is more important than the ability to speak in learning any language. At the stage of 5 to 6 years, the vocabulary of child is so developed that, he can comprehend the meanings of stories told to him.. According to Hurlock gestures help in understanding the meaning more properly as compared with mere words. Multimedia provides a facility to use animated graphics, sound effects, music etc. It is sure that the children can understand the stories in more details.

**Vocabulary building** : The child's vocabulary level is connected with intelligence and motivation. When multimedia is used, it is not only conducive to the learning but also to find new meanings of the words already learned. The use of radio, television and films in the schools will also help the increase of vocabulary. With the advent of multimedia one can design interactive games which will help in building vocabulary. The multimedia games also initiates group discussion and debates amongst the children of this group.

**Formation of sentences** : Children begin to construct simple sentences involving two or more words and slowly grow to more words and begin to construct compound sentences. The story telling with the help of multimedia, is an effective tool in motivating the children in sentence formation.

**Correct pronunciation** : Children learn to pronunce words by imitation of others. The occasions of training and the amount of motivation is the main source behind the level of pronunciation.

### *Supplimentary materials and their needs*

In any formal language learning / teaching situation, there is a necessity to have supplementary materials with a view to have :

(1) Re-inforcement of language learning

(2) Coverage of certain aspects which have not otherwise covered in the conventional material

(3) To supply additional information / teaching inputs.

(4) To overcome certain deficiencies in teaching as well as materials.

Supplementary materials are of different types such as print, audio, video computer based training, multimedia, internet etc. Each kind has its advantages as well as disadvantages. Also, the material preparation for different media varies. For example the material prepared for print media may not be applicable for the audio or video media.

In this paper an attempt has been made to bring out the usefulness of the multimedia towards learning of a school language which is diglossic in nature. It may be mentioned that the material preparation for this multimedia concentrates towards the learning of school language keeping in mind the linguistic difference between the home dialect and school language (written Tamil). In other words, it may be mentioned that, spoken Tamil material is not utilised from which switch over stragety could have been attempted to reach written Tamil for the various social linguistic reasons. But, having kept in mind the linguistic differences between Home dialect and school language at the phonological, grammatical, lexical and semantic levels, materials in school language have been prepared and provided for the multimedia.

### *Multimedia*

Multimedia consists of Text, graphics, video, animation, sound and voice. Multimedia technology when used in creating supplementary material emphasises and enhances language learning in an interactive way using the live situation, which is not possible in any other media.

### *Sentamil*

Sentamil is a multimedia based software for learning Tamil (which could be made for any language) as a first language for those who have Tamil as their mother tongue (home dialect).

The main objective of Sentamil development is to develop, re-inforce and enhance all the four basic language learning skills which may generally be learned in a conventional class room teaching/learning situation. The presentation of materials are such that, the other objective of Sentamil is to present the material for effective learning through entertainment.

Sentamil is comprised of the following components :

(1) Moral stories / narratives for developing listening

(2) Words and sentences for development of vocabulary / reading

(3) Words for spelling / pronunciation practice

(4) Script teaching with hand movements.

**Listening Skills** : Simple moral stories and other stories which can inculcate the Tamil culture and ecosystem are presented in its natural environment. The stories are presented with suitable visuals and necessary animations and narrated with dramatization and description. The narration is in simple language with clear pronunciation, intonation and expression. The development of listenting skills is evaluated through newly designed and interesting exercises, The Development of sequencing in presentation through coherance and cohessiveness are also tested.

**Vocabulary teaching** : Graded vocabulary with its proper pronunciations is given to increase the learner's lexicon. Word building and word formation are explained in such a way that the learners can interestingly interact with the package and learn. Ample exercises and innovative and interactive language games are included to improve the vocabulary development.

**Speaking** : Designed in such a way that the learner can acquire the speaking skill with proper pronunciation and intonaion. Provision for group drills to be monitered by the teacher is also given.

**Script teaching and Writing** : Introduces all Tamil letters through shape similarity and pattern perception method as adopted in the Tamil Nadu text books. It emphasizes the appropriate hand movements, with animation, sound symbol correlation, letter in words, words with proper pronunciation etc are the major characterstic features. Reinforecement of script learning through exercises. Learner centered provision for error correction are also given.

Provision is made to acquire the appropriate Tamil orthographic strokes, in letter, word and sentence levels. This is done by way of giving appropriate instruction given in the package for writing practice. Facility to practice writing after listening to the pronunciation and also by viewing the text in all levels is also provided.

**Exercise and games** : The package comprises of a number of creative and interactive exercises and games. which are highly suitable to the computer media. In a sense they demand active participation of the learners, monitor the children in language learning and create group interaction. In addition to these, they develop the analytical and creative thinking of the children.

### *Material preparation for Sentamil.*

Material preparation for multimedia based contents is a new experience as there was no materials / softwares available for syllabus based learning. I myself got involved during the material preparation. During preparation, it was learnt that more care was to be taken for ensuring the multimedia from all angles.

### *Scope of Sentamil*

Sentamil is first of its kind for any of the Indian languages. With lot of exercises and games, this type of material could be prepared for other Indian languages as well. This product is not only useful as suplementary but could also be used as stand alone aid. The children can use the CD on their own pace and learn Tamil leisurely.

The multimedia based Teaching / learning materials (Aids) bring life and helps pupils better understand the abstract difficult concepts easier and grasp. It also motivates the learning interest of the pupils with the real life animations with sound effects and music in it and helps reduce the labour of the teachers. Multimedia also enables the learners in re-inforcement the subject by means of games and in turn paves ways for debates and discussions. Material development are influenced by these and help gross human development.

### *Conclusion*

Experts say that there are four Es (Explore, Experience, Enlarge and Evaluation) to take care of proper development of education methods. Training aid in the form of multimedia language learning software comes handy with all these perceptions and again the package is conceived, devised and produced with all these perceptions to serve the School language to cater for the blossoming buds and spread of its strength of augmenting the existing educational training aids, less this fascinating method.

## Colophon

I wish to thank Dr V. Gnanasundaran, Dr N. Nadaraja Pillai, Dr Sam Mohan Lal and Dr M.Ganesan, all from CIIL Mysore, who are involved with me in the preparation of Sentamil.

## BIBLIOGRAPHY

1. TAMIL - An Auto Insructional Course by Dr N. Deivasundarm and Dr. A. Gopal, International Institute of Tamil Studies
2. Hand Book TAMIL - An Auto Instructional Course by Dr S. Jean Lawrence and Dr. D.Renganathan of International Institute of Tamil Studies
3. A Syntactic Study of TAMIL Verbs by Shri (Now Dr) N. Nadaraja Pillai of CIIL
4. A Linguistic Study of the Problems in Learning Tamil by Tribal children of the Nilgiris by Shri C. SIVAKUMAR of CIIL
5. A Programmed Course in TAMIL by Shri G. SANKARANARAYANAN of CIIL
6. Social Psychological Perspectives on Second Language Learning by Shri R.K.Agnihotri, A.L. Khanna and I. Sachdev for SAGE Publications.
7. School Psychology by Dr Ramnath Sharma for RAJHANS Publications
8. Sentamil by APPLESOFT, Bangalore - 560 010.

# தமிழ் ஆராய்ச்சிக்குக் கைகொடுக்கும் தகவல் தொழில்நுட்பம்

சு. சீனிவாசன், அறிவியல் அலுவலர்,
கணிப்பொறிக் கோட்டம்,
இந்திராகாந்தி அணுவாராய்ச்சி மையம்,
கல்பாக்கம். 603 102

தமிழில் கருத்துப் பரிமாற்றம் பேச்சு மூலமும் எழுத்து வடிவத்திலும் பல நூற்றாண்டுகளாக நடைபெற்று வருகிறது. பேச்சுவழி செய்திகளைப் பதிவு செய்யும் முயற்சி இருபதாம் நூற்றாண்டில் தான் தொடங்கிற்று. எழுத்துவடிவ பரிமாற்றம் பல நூற்றாண்டுகளாகப் பதிவு செய்யப்பட்டுக் கிடைக்கிறது. தமிழின் தொன்மைக் கால இலக்கியங்கள் பலவும் செய்யுள் வடிவிலேயே இருக்கின்றன. கடந்த சில நூற்றாண்டுகளாக உரைநடையில் எழுதும் போக்கு வளர்ந்திருக்கிறது.

அண்மைக் காலங்களில் மொழியின் பண்பைத் தகவற் கோட்பாட்டின் (Information Theory) அடிப்படையில் கணிக்கும் வழிமுறைகள் தோன்றி இருக்கின்றன. தொலை இடங்களுக்குத் தகவலை உரையாகவோ (Text), வரைவாகவோ (Graphics), பேச்சாகவோ (Speech) விரைந்து அனுப்ப மொழியின் பண்புகளை ஆய்வுசெய்யும் நெறிமுறைகள் வகுக்கப்பட்டுள்ளன. இப்பண்புகள் மொழிகளுக்கிடையே வேறுபடுகின்றன.

கணிப்பொறி (Computer), தொடர்பு (Communication), வலையமைப்பு (Network) ஆகியவற்றின் சங்கமமாகக் கருதப்படும் தகவல்நுட்பத் துறையின் வளர்ச்சி அசுர வேகத்தில் நிகழ்ந்து வருகிறது. அறிவியல் வளர்ச்சியில் கணிப்பொறி ஒரு கருவியாக மட்டுமல்லாமல், ஆய்வுக் கூடமாகவும் (Laboratory) விளங்குகிறது. பல பரிமாணத் தோற்றங்களை எளிதில் கணிப்பொறித் திரையில் கொண்டுவந்து விளங்கிக்கொள்ள 'வரைகலை' (Graphics) பேருதவியாக இருக்கிறது. உலக வலைப்பின்னல் (World Wide Web) எனப்படும் இணையத்தின் வாயிலாக எத்துறைப் பற்றிய செய்தியையும் தகவலையும் பெற முடிகிறது. வலிமையான தேடல் பொறிகள் (Search engines) இலக்கியங்களைச் சில மணித்துளிகளில் துருவி ஆராய்ந்து நாம் கேட்கும் தகவல்களைத் திரைமுன் கொண்டுவந்து கொடுக்கின்றன. ஒளி–ஒலி தகவல்களைக் கோப்புகளாக (Files) உடனுக்குடன் பரிமாறிக்கொள்ள முடிகிறது. படக்காட்சி முறையில் தொலைமாநாடுகளையும் நடத்த இயலுகின்றது. மின்–அஞ்சல் போன்று மின்–இதழ்களும் (E-Journals) இணையத்தில் (Internet) வாசிக்கக் கிடைக்கின்றன.

தையற்பொறி, கடைசல் எந்திரம் போல கணிப்பொறியும் ஒரு கருவியே. இருப்பினும் இஃது ஒரு குறிப்பிட்ட பணிக்கென அல்லாமல் பல் துறைகளிலும் பயன்படவல்லது. ஈராயிரம் ஆண்டு பழமை உடைத்த தமிழ் மொழியில் அமைந்துள்ள 'யாப்பிலக்கணக் கோட்பாடு' கணித அடிப்படையில் வேரூன்றி மிளிர்வதைக் காணலாம். இரும இலக்கம் (Binary Digit) என்று நாம் கூறும் உண்டு/இல்லை பண்பை நேர்/நிரை என்ற 'அசை' உறுப்பாக நம் முன்னோர்கள் கையாண்டிருக்கிறார்கள். எழுத்தினாலாவது அசை; அசையினாலாவது சீர்; பல சீர்களால் அமைவது அடி–என்பதற்கு ஒப்பாக கணிப்பொறியில் சேமிக்கப்படும் தகவல்கள் bit (the smallest unit of storage), byte (fixed number of bits), word (collection of bytes) என வகைப்படுத்தப் பட்டிருக்கின்றன.

தகவல்களைச் செலுத்துவதற்காகும் நேரத்தை எங்ஙனம் திறம்படக் குறைப்பது என்பதைக் காண இன்றைய தகவல் கொள்கை முயலுகிறது. பண்டைய சமுதாயம் இத்தகு தகவல்களை எவ்வாறு எளிதாகவும் விரைவாகவும் ஊடகத்தில் பதிவு செய்யலாம் என்பதை நோக்கமாகக் கொண்டிருந்தது. தமிழ் உரைநடையில் பயிலும் உயிர்மெய் எழுத்துகள் இன்றைய தகவல்நுட்பத்திற்கு உகந்த முறையில் அமைந்துள்ளன. அன்று ஏட்டில் பதிவு செய்வதற்குக் காட்டிய முக்கியத்துவத்தை, இன்று நாம் தகவல் தொடர்பில் புகுத்திக் கொண்டிருக்கிறோம்.

தமிழ் எழுத்து வடிவம் ஆங்கிலத்தைப் போல் சமநீளக் குறியீடாக (Fixed length code) அமையாமல், மாறும் நீளம் கொண்ட குறுயீடாக (Variable length code) அமைந்திருக்கிறது. இதன் வரலாற்றுப் பின்னணியில் தகவல் செறிவு (data compression) என்ற பண்பு மறைந்திருப்பதை இன்றைய தகவற் கோட்பாட்டைக் கொண்டு விளக்க முடிகிறது. தந்திமுறையில்

தகவல் அனுப்பப் பயன்படும் மோர்சு குறிமுறையில், அதிகமாக ஆளப்படும் எழுத்துகளுக்கு நீண்ட குறியீடுகளும் அமைக்கப்பட்டுள்ளன. மோர்சு குறிமுறை அடிப்படையில், தமிழ் நெடுங்கணக்கை ஆராய்வோமானால் ஓர் உண்மை புலப்படுவதைக் காணலாம். இங்கு ஒரெழுத்து வடிவிலான அகர, உகர உயிர்மெய் வரிசை எழுத்துகள் அதிகம் புழங்குவதையும், மூன்றெழுத்து வடிவிலான ஒகர, ஓகார உயிர்மெய் வரிசை எழுத்துகள் அரிதில் புழங்குவதையும் மதிப்பிட முடிகிறது. நவீன தகவல்நுட்பக் கோட்பாட்டுக்கு ஒத்த வகையில் தமிழ் உயிர்மெய் எழுத்துகள் மெல்லப் படிமலர்ச்சி (evolution) கண்டு முகிழ்த்திருக்கின்றன என்ற செய்தி வியப்புக்குரியது.

இத்தகு சிறப்புமிக்க உயிர்மெய் எழுத்துகளை, கணிப்பொறியில் உள்ளிடவும் பதிவு செய்யவும் நேர்த்தியான உத்திகளைப் பின்பற்றுகிறோமா என்று வினா எழுப்பின் குழப்பமே மிஞ்சுகிறது. வகுப்பறைக்கு வரும் மாணவர்கள் எந்த வண்ண ஆடையும் அணிந்து வரலாம் என்று சுதந்திரம் அளிக்கலாமா? வகுப்பறைக்கு வந்தபின் அங்குள்ள இருக்கையில் எங்கு வேண்டுமானாலும் அமரலாம் என்று அனுமதிக்கலாமா? இவற்றுக்கு விடை கண்டுபிடித்துவிட்டால், கணிப்பொறியிலும் தமிழ் எழுத்துகள் களிநடம்புரியத் தொடங்கிவிடும் என்று கேட்க உங்களுக்கு வியப்பாக இருக்கலாம். இதற்கு உகந்த சாவிப்பலகை அமைப்பும் (keyboard layout) குறியீடுகளும் (codes) நெறிப்படுத்தப்பட்ட வேண்டும். கலங்கிய நீர் தெளியட்டும் எனக் காத்திராமல், குண்டானில் படிகாரத்தைப் போட்டு நீரை இறுத்துப் பருக வேண்டிய நேரம் வந்தாயிற்று.

கணிப்பொறித் துறையில் நாள்தோறும் ஏற்பட்டுவரும் வளர்ச்சி காரணமாக, இன்று பொறிகளுக்கு கட்புலன், செவிப்புலன், பேச்சுத் திறன் ஆகியவற்றை ஊட்டும் முயற்சிகள் உலகின் பல மூலைகளில் நடைபெற்று வருகின்றன. இப்பணிகள் அனைத்திற்கும் அடிப்படையில் தேவைப்படுவது தகவலை இலக்கமாக்கும் (digitising) வழிமுறையாகும். எழுத்திலிருந்து பேச்சு உருவாக்கம் (text-to-speech), கையெழுத்தை அடையாளம் காணல் (hand written character recognition), பேச்சை அறிதல் (speech recognition), பேசுவோரை இனம்காணல் (speaker verification) ஆகியவற்றுக்கு இன்று கணிப்பொறியுடன் கூடிய மென்பொருள்கள் (softwares) உருவாக்கப்பட்டிருக்கின்றன. இவற்றில் தமிழில் அச்சேறிய எழுத்துகள் அடையாளம் காணும் முறை தரவுத்தள (database) பணிகளை விரைந்து மேற்கொள்ள உதவிடும்.

தமிழ் எழுத்துகளின் வரைவியல் குறித்த வருணனை (geometrical description of characters) தமிழ் இலக்கண நூல்களில் விரிவாக கூறப்படவில்லை. அரசுப் பள்ளிகளில் படிக்கும் குழந்தைகள், இரண்டாம் வகுப்பிலேயே அனைத்து தமிழ் வரிவடிவுகளையும் எழுதக் கற்றுக்கொண்டு விடுகின்றனர். இது தொடர்ந்த பயிற்சியின் காரணமாகவே சாத்தியமாகிறது. அவர்கள், எழுத்துகளை விரைந்து அடையாளம் கண்டு வாசிக்கவும் பயிற்சி பெற்றுவிடுகிறார்கள். இந்தக் கற்றல் நிகழ்வின் (learning phase) பின்னணியில் இன்றைய தமிழ் எழுத்து வடிவின் பல்வேறு கூறுகளை, நாம் நினைவில் நிறுத்துகிறோம். அவற்றில் பெரும்பாலானவற்றை ‘வரைவழி நாற்பது’ எனும் தலைப்பில் தொகுத்துள்ளோம். இவற்றில் சில படம் 1ல் கொடுக்கப்பட்டுள்ளன.

மேலே கொடுக்கப்பட்டவை போன்ற வரிவடிவின் பண்புகள் நாற்பதையும் உணர்வு நிலையில் கருவிக்கு ஊட்டுவது எளிய பணியன்று. இச் செயலுக்கு ‘செயற்கை நரம்பணு வலைப்பின்னல்’ (artificial neural network) என்ற வழிமுறை பின்பற்றப்படுகிறது. சிக்கல் வாய்ந்த ‘கையெழுத்தை அடையைாளம் காணும் பணி’ முதலியவற்றுக்கு இந்த வழிமுறை உகந்தது. தமிழ்ல் அச்சேறிய எழுத்துகளை பகுத்து அடையாளம் காண வரைவியல் (geometry) அடிப்படையிலான செயல்முறை ஏற்புடையது. இப் பணிக்கு ‘ஒளிசார் எழுத்துரு அடையாளம் காணல் முறை’ (optical character recognition method) செயல்வடிவம் அளிக்கின்றது.

திராவிட மொழிகளில் குறைந்த எழுத்துருவங்களைக் கொண்ட மொழி தமிழ் ஒன்றே! தமிழ் உரைநடையில் பயிலும் சொற்களின் சராசரி நீளம் 4.4 எழுத்துகள். அதாவது பெரும்பாலான சொற்கள் 4 முதல் 5 எழுத்து நீளமுள்ளவையாக இருக்கின்றன மலையாளத்தில் சொற்களின் சராசரி நீளம் 5.9. அதாவது பெருவாரியான சொற்கள் 5 முதல் 6 எழுத்து நீளமுடையன. வடவெழுத்துகள் அதிகம் புழங்கப்படும் மலையாளத்தில் சந்திகளைக்கொண்டு சொற்களை இணைக்கும் பண்பு மிகுதி. மொழியின் நெடுங்கணக்கை வல்லினம், இடையினம், மெல்லினம் எனப் பிரித்து, இனவெழுத்துகளின் புழக்கமும் கணிக்கப்பட்டது. இனவெழுத்துகளில் சுமார் 50 விழுக்காடு வல்லெழுத்துகளே அரசோச்சுகின்றன. தமிழ், மலையாள மொழிகளில் இடையினமும் மெல்லினமும் 1 : 1 என்ற விகிதத்தில் கையாளப்படுகின்றன. கன்னடம், தெலுங்கு மொழிகளில் இவ்விகிதம் 2 : 1 என்ற அளவில் அமைகிறது. தமிழில் 18 மெய்வரிசை எழுத்துகள் இருப்பினும்,

பண்பு 1: சிறகுகள் குவித்த எழுத்துருக்களாவன:

ஈ, உ, ஊ, எ, க, கூ, ங, ச, சு, ட, டீ, ண, ப, ம, ய, ல, வ, ள, ன, ஸ, ா, ை, ௌ (23)

பண்பு 2 : சிறகுகள் விரித்த எழுத்துருக்களாவன:

இ, ஞ், ஞி, ஞீ, டு, டூ, த், தி, தீ, ந், நி, நீ, ழ், ழி, ழீ, ற், றி, றீ, ஜ், ஜி, ஜீ, ஷ், ஷி, ஷீ, ஹ், ஹி, ஹீ, க்ஷ், க்ஷி, க்ஷீ, ஸ்ரீ (31)

பண்பு 3: மேற்புறம் சிறகு விரித்த எழுத்துருக்களாவன:

ஃ, க், கி, கீ, ங், ஙி, ஙீ, ச், சி, சீ, ட், ண், ணி, ணீ, ப், பி, பீ, ம், மி, மீ, ய், யி, யீ, ர், ரி, ரீ, ல், லி, லீ, வ், வி, வீ, ள், ளி, ளீ, ன், னி, னீ, ஸ், ஸி, ஸீ, ெ, ே, ௌ (44)

பண்பு 4: கீழ்புறம் சிறகு விரித்த எழுத்துருக்களாவன:

அ, ஆ, ஏ, ஐ, ஒ, ஓ, கு, ஙு, நூ, சூ, ஞ, ஞு, ஞூ, டி, ணு, ணூ த, து, தூ, ந, நு, நூ, பு, பூ, மு, மூ, யு, யூ, ர, ரு, ரூ, லு, லூ, வு, வூ, ழ, ழு, ழூ, ளு, ளூ, ற, று, றூ, னு, னூ, ஜ, ஷ, ஹ, க்ஷ (49)

பண்பு 5: சுழியில் தொடங்கும் எழுத்துருக்களாவன:

அ, ஆ, இ, உ, ஊ, எ, ஏ, ஐ, ஒ, ஓ, ஞ், ஞ, ஞி, ஞீ, ஞு, ஞூ, ண், ண, ணி, ணீ, ணு, ணூ, ல், ல, லி, லீ, லு, லூ, வ், வ, வி, வீ, வு, வூ, ள், ள, ளி, ளீ, ளு, ளூ, ன், ன, னி, னீ, னு, னூ, ஜ், ஜ, ஜி, ஜீ, ஷ், ஷ, ஷி, ஷீ, ஸ், ஸ, ஸி, ஸீ, ஹ் ஹ, ஹி, ஹீ, ெ, ே, ை (65)

பண்பு 6: சுழியில் முடியும் எழுத்துருக்களாவன:

ஓ, கீ, ஙீ, சீ, ஞீ, டீ, டூ, ணீ, தீ, நீ, பீ, மீ, மூ, யீ, ரீ, ரூ, லீ, வீ, ழீ, மூ, ளீ, ளூ, றீ, னீ, ஜீ, ஷீ, ஸீ, ஹீ, க்ஷீ, ஸ்ரீ, ே (31)

பண்பு 7: சுழியில் தொடங்கி சுழியில் முடியும் எழுத்துருக்களாவன:

ஓ, ஞீ, ணீ, லீ, வீ, ளீ, ளூ, னீ, ஜீ, ஷீ, ஸீ, ஹீ, ே (13)

படம் 1 வரைவழி நாற்பது

இவற்றில் அனைத்து வரிசை எழுத்துகள் இருப்பினும், இவற்றில் அனைத்து வரிசை எழுத்துகளும் சமஅளவில் பயனாவதில்லை. எடுகொட்டாக ங, ஞ வரிசை எழுத்துகள் அரிதே வருகின்றன. கணிப்பொறிக் கொண்டு உரைநடையை ஆராய்ந்தபோது தமிழில் பயனுடைய மெய்வரிசை எண்ணிக்கை 14 என நிறுவ முடிந்திருக்கிறது.

ஈராயிரம் ஆண்டுகளுக்கு மேலாக வழக்குடையதாக விளங்கும் தமிழ் மொழியில் உள்ள நூற்பகுதிகள் சில ஒரே ஆசிரியரால் எழுதப்பட்டனவா அல்லது அவற்றின் ஆக்கத்தில் பலர் தொடர்புடையவர்களாக இருந்தார்களா என்ற ஐயம் அறிஞர் பலருக்கு எழுவது உண்டு. நூற்பகுதியில் காணும் நடைவேறுபாட்டைக் கொண்டு மொழியியலார் சிலர், இங்ஙனம் முடிவுகட்டுவது இயல்பே. இக்கூற்றுக்கு முழுமையான விடைகாண்பது இயலாது என்ற போதிலும், நுலாசிரியரை அடையாளம் காண எழுத்துறவுக் கணிப்புமுறையைப் (letter correlation computing method) பயன்படுத்த முடிகிறது. திருவள்ளுவர், கம்பர், பாரதியார் ஆகியோரின் நடை அவர்தம்மையின்றி பிற நுலாசிரியரோடு ஒத்துப் போவதில்லை என்பதைக் கணிப்பிலிருந்து அறிகிறோம். ஒரு உரையில் காணப்படும் எழுத்துகளின் நிகழ்வு (occurrence), எழுத்துச் சரங்களின் உறவு (letter string correlation), அடுத்துவரும் சொற்களுக்கிடையேயான ஒழுக்கு

(order), சொற்களின் புழக்கம் (frequency), செய்யுள் நடையில் பயிலும் யாப்பு இவற்றைப் புள்ளியியல் அடிப்படையில் ஆய்வு செய்வது பயன் அளிக்கக் கூடியது.

இனவெழுத்துகளின் புழக்கம் (க - ங, ச - ஞ, ட - ண, த - ந, ப - ம, ற - ன), வெண்பா யாப்பில் பயிலும் தளை (மா முன் நிரை, வின முன் நேர், காய் முன் நேர்), வாக்கியங்களில் தொடர்ந்து வரும் வேற்றுமை உருபுகளின் (ஐ, ஆல், கு, இன், அது, கண்) பயன்பாடு ஆகியவற்றை மார்க்கோவ் செயல்முறை கொண்டு ஆராய முடிகிறது. இப்பணியை விரைந்து மேற்கொள்ள, கணிப்பொறி கைகொடுக்கிறது. விண்ணில் ஏவுகளை ஒன்றைச் செலுத்த நாள் ஒன்றைத் தெரிவுசெய்ய வேண்டியிருக்கிறது எனக் கொள்வோம். தொடர்ந்து 7 நாட்கள் 'வெயில்' பொழுது இருக்கவேண்டும் என்ற நிபந்தனைப் பிறப்பித்து நாள் குறிப்பதற்கு 'மார்க்கோவ்' செயல்முறை பயன்படுகிறது. இதையொத்த கேள்வி ஒன்றை திருக்குறள் நூலிலும் எழுப்பிடலாம். குறள் வெண்பாவில், அனைத்தும் காய்ச்சீர் கொண்ட குறள்கள் எத்தனை வருகின்றன? வள்ளுவர் நடையில் – முழுவதும் வெண்சீர் வெண்டளை கொண்ட குறளின் விழுக்காடு என்ன? ஆகியவற்றுக்கு 'மார்க்கோவ்' செயல்முறை பதிலளிக்கிறது. வள்ளுவர் நடையில் ஏறக்குறைய 5000 குறள்களுக்கு ஒரு முறையே காய்ச்சீர் தொடர்ந்து வரக்கூடும் என்பது முடிவாக இருப்பின், வள்ளுவர் தாம் யாத்த நுலைப் போன்று மும்மடங்கு எழுதினாலும்,

"யாதானும் நாடாமால் ஊராமால் என்னொருவர் சாந்துணையுங் கல்லாத வாறு. (397)"

எனும் முற்றிலும் காய்ச்சீர் கொண்ட குறள் போன்று மீண்டும் ஒன்று வாராது என்பதை நிறுவ முடிகிறது. நாலடியார் நூலை பலர் எழுதினார்களா என்பதற்கும் தளை ஆய்வு உதவிடுகிறது.

சிறுகதைகளின் நடைப் பண்பை மதிப்பட வாக்கிய நீளம் ஒரு காரணியாக அமைகிறது. வருணனை மிகும் இடங்களையும், உணர்வு வெளிப்படும் இடங்களையும் வாக்கியத்தில் இடம்பெறும் சொற்களின் எண்ணிக்கையிலிருந்து ஒருவாறு கணிக்க முடிகிறது. சிறுவர்களுக்கான படைப்பு இலக்கியங்களில் எளிய வாக்கிய அமைப்பு கொண்ட நடையை தெரிவுசெய்ய வாக்கிய நீள மதிப்பீடு ஓர் அலகாக உதவிடுகிறது. அகராதி உருவாக்கும் பணியில் புதுமைகள் பல நிகழ்ந்த வண்ணம் இருக்கின்றன. அன்றாட வாழ்வை வழிநடத்திச் செல்லவும் இடம்பெயரும் புதிய மக்கள் சூழலுக்கு ஏற்ப, அடிப்படை வழக்குச் சொற்களைத் (vocabulary) தெரிந்துகொள்ளவும் முயற்சிகள் மேற்கொள்ளப்படுகின்றன. புதிதாக மொழியைக் கற்போருக்கு இத்தகு சொற்களைத் திரட்டித்தருவது பயன் அளிக்கக்கூடியது. நாளேடுகளில் வரும் தலைப்புச் செய்திகளை புரிந்துகொள்ளும் வகையில் இச்சொல் திரட்டு அமைவது நன்று. சிக்கல் மிகுந்த எழுத்துகள் கொண்ட சீன, ஜப்பான் மொழிகளைக் கற்க, சொல் புழக்கத்தின் அடிப்படையில் அமைந்த அகராதிகள் உதவுகின்றன. அவற்றை ஒத்த முயற்சிகள் கணிப்பொறி துணை கொண்டு தமிழிலும் மேற்கொள்ளப்படல் வேண்டும்.

பிறமொழியிலுள்ள அறிவுக் கருவூலங்களை நுகர்வதற்கு மொழி பெயர்ப்பு இன்றியமையா ததாகிறது. எல்லா அறிவுச் செல்வங்களும் ஒரே மொழியில் கிடைத்துவிடுவதில்லை. பொதுவாக, மொழிபெயர்ப்பு என்பது செம்மையாகவும் நம்பகத்தன்மையுடனும் இருத்தல் வேண்டும். மொழியின் தன்மைக்கேற்ப, மொழிபெயர்க்கப்பட்ட பகுதியின் அளவு வீக்கமுறவோ சுருங்கவோ செய்கிறது. இவற்றை அளந்து மதிப்பிட வழிமுறை காண்பது அவசியமாகிறது. உரையின் வரி நீளம், தகவற் திணிவு (information content or entropy) ஆகிய அளவு கோலை அடிப்படையாகக் கொண்டு மொழிபெயர்ப்பை மதிப்பிட முடிகிறது.

இந்திய மொழிகளில், குறிப்பாக தமிழ்மொழிக்கு இரு சிறப்புகள் உண்டு. முதல் சிறப்பு தமிழ்நெடுங்கணக்கில் எழுத்துகள் குறைவு என்பதும் கூட்டெழுத்துகள் இல்லை என்பதாகும். இரண்டாவது சிறப்பு, ஒரே வல்லெழுத்து பல நிலைகளில் மாறுபட்ட ஒலிகளை உருவாக்குவதற்கு தமிழில் நெறிகள் வகுக்கப்பட்டிருப்பதாகும். ஐரோப்பிய மொழிகளின் எழுத்தமைப்பு எளிது என்றாலும் அவற்றிலுள்ள சொற்களை உச்சரிப்பதற்கு மிகுந்த பட்டறிவு தேவைப்படுகிறது. நம்மொழிகளில் எழுத்தமைப்பு சிக்கலானது என்றாலும் எழுதுவதைப் போலவே உச்சரிக்கிறோம். எனவே உரையிலிருந்து பேச்சு உருவாக்கம் என்பது தமிழில் எளிதாக மேற்கொள்ளக்கூடிய பணியாகும்.

பேச்சுத் தமிழைப் பதிவுசெய்து அவற்றினின்று, குறில், நெடில் எழுத்துகளுக்கான மாத்திரை அளவை துல்லியமாக கணக்கிடமுடிகிறது. குறில் எழுத்துகளுக்குரிய ஒலிப்பதிவு நேரம் 150 மில்லி செகண்டுக்கு உள்ளாகவும் நெடில் எழுத்துகளுக்கான நேரம் 200 மில்லி செகண்டுக்கு

உள்ளாகவும் இருக்கின்றன. வல்லொற்றுகளை பதிவு செய்ய கூடுதல் கவனம் தேவைப்படுகிறது. இவ்வொலிப் பதிவில் அமைதிப் பள்ளத்தாக்குகள் (silence period) இருப்பது அவைநேர்த்தி செய்யும்போது வெளிப்படுகிறது. தமிழிசைப் பண்களையும் தாளங்களையும் விரிவாக ஆராயவும் கணிப்பொறி கைகொடுக்கிறது.

கணிப்பொறித் தொழில்நுட்பம், தமிழ் ஆராய்ச்சியின் பரிமாணத்தை விரிவுபடுத்தவும் விரைவுபடுத்தவும் உதவிடக் கூடும். இருப்பினும் இத்தொழில்நுட்பம் கடைக்கோடி மக்களுக்கும் கிட்டவேண்டுமென்றால், இத் துறைசார்ந்த நூல்கள் தமிழில் வெளிவருதல் அவசியம். அண்மையில், அண்ணாப் பல்கலைக்கழகம் கணிப்பொறிக் கலைச்சொல் அகராதி ஒன்றை தமிழில் வெளியிட்டு இருக்கிறது. இதில் நாலாயிரத்திற்கும் மேற்பட்ட கலைச்சொற்கள் இடம்பெற்றுள்ளன. தமிழ் மொழிக்கான சாவிப்பலகை அமைப்பும் குறியீடுகளும் தரப்படுத்தப்பட்டு விட்டால், நம் பயணம் எளிதாகும். அகராதிகளையும், தரவுத் தளங்களையும் தமிழ் வாயிலாக கணிப்பொறியில் நிறுவி, இணையத்தின் மூலம் தொடர்புகொள்ளும் நாட்கள் வெகுதொலைவில் இல்லை என்றே எண்ணத் தோன்றுகிறது. வாழ்க தமிழ்; வளர்க தமிழர் பணி!

**18.01.99**

# தொன்ம வேர்களை இணைக்கும் மின் வேர்!

சா.அ.சௌரிராசன்
தொழில்நுட்ப இயக்குநர், இன்தாம் இணையம்,
1-பி, ஜே.பி.டவர் முதல் தளம், 7/2, நுங்கம்பாக்கம் நெடுஞ்சாலை, சென்னை - 600 034,
தொலைபேசி : 8275639, 8236473, மின்னஞ்சல் : sowri@intamm.com,
இணைய பரப்புகள் : www.intamm.com, www.tamilisai.com, www.thirukkural.com

## தமிழர்கள் நிமிர்ந்ததும் நிலை குலைந்ததும்

உலகின் ஆறு கண்டங்களிலும், அறுபத்தைந்து நாடுகளுக்கு மேற்பட்டு, ஒன்பது கோடித் தமிழர்கள் வாழ்ந்து கொண்டிருக்கிறார்கள்.

உலக நாடுகளில் எங்கும் பரவியிருக்கும் தமிழ் இனம், ஏன் பரவியது? எப்படிப் பரவியது? சங்க காலத்தில் திரைகடல் ஓடி திரவியம் தேடிய தமிழர்கள், 10 ஆம் நூற்றாண்டில் கீழ்த்திசை நாடுகளை வெற்றி கொண்டு அரசாண்ட தமிழர்கள், 17 ஆம் நூற்றாண்டின் இறுதியிலும், 18 & 19 ஆம் நூற்றாண்டுகளிலும் ஆங்கிலேயர்களுக்கு அடிமைச் சேவகம் செய்ய கூலிகளாக உலகம் முழுவதும் பரவினர்.

தமிழர்கள், மலேயாவின் பால் மரக் காடுகளிலும், இலங்கையின் தேயிலைத் தோட்டங்களிலும், மொரிஷியஸ், பிஜி, பிரஞ்ச் கயானா, பிரிட்டிஷ் கயானா, சுரிநாம், மேற்கிந்தியத் தீவுகள், போன்ற நாடுகளின் கரும்புத் தோட்டங்களிலும், தென்னாப்பிரிக்காவின் சுரங்கங்களிலும் ரத்தத்தை வியர்வையாக்கி உழைத்து அட்டைகளுக்கும், மலேரியா கொசுக்களுக்கும், விபத்துக்களுக்கும் தம் இன்னுயிரை இழந்தனர். எஞ்சியோர், வாழ்க்கையின் கடைக் கோடி நிலையில் போராடி உழைப்பு, உழைப்பு, என்பதைத் தவிர உயர்வு உயர்வு என்பதை அறியாமலேயே தத்தமது நாடுகளின் பொருளாதார வளத்திற்கு அடித்தளமிட்டு உயர்த்தினர்.

## தொடர்பின்மையே

அத்தகைய தமிழர்களுக்கு வரலாறு இட்ட பெயர்கள் கூலி, அடிமை, மாவுச் சாக்கு, கோணி, சாலா மதராஸி, கட்டா பாணி, கள்ளத் தோணி, நாடற்றவர், அகதி என்றெல்லாம் நீள்கிறது. ஆனால் இதே வரலாற்றில் இங்கிலாந்து நாட்டில் உற்பத்தி செய்யப்பட்ட பொருட்களை விற்பனை செய்வதற்காக புறப்பட்ட ஆங்கிலேயர்கள், வணிகம் செய்து அதன் தொடர்ச்சியாக நாடுகளையும் கைப்பற்றி ஆட்சி செய்து தமது ஆங்கில மொழியையும் பரவச் செய்து நிலை பெறவைத்தனர். அடிமையாய் கூலியாய் புறப்பட்ட தமிழன் தனது தாய்மொழியை மறந்தான், பண்பாட்டுக் கூறுகளை இழந்தான், வேற்று மொழியோடும் இனத்தோடும் கலந்தான். இன்று பெயர்களிலும், சில மதச் சடங்குகளிலும் மட்டுமே அடையாளம் காணப்படக் கூடியவனாக வாழ்ந்து கொண்டிருக்கிறான். இதற்குக் காரணம் என்ன? தொடர்பின்மையே! சென்ற நூற்றாண்டுளில் தாய்த் தமிழகத்திற்கும், மற்ற நாடுகளில் பரவிய தமிழர்களுக்கும் தொடர்ச்சியான தொடர்பு இல்லாமல் இருந்ததுதான் அடிப்படை காரணமாகும்.

## எழுத்துத் தமிழை இழந்தது

உலகம் பரவிய தமிழர்களின் முதல் தலைமுறை தமது தாய் மொழியாகிய தமிழை பேச்சு, எழுத்து என இருவடிவங்களிலும் பயன்படுத்தியது. ஆனால் அடுத்த தலைமுறை எழுத்தை இழந்தது. அதற்கடுத்த தலைமுறை பேச்சுத் தமிழைப் புரிந்து கொள்ளக்கூடிய அளவிற்கு மாறியது. அதற்கடுத்த தலைமுறை பெயரில் மட்டுமே அடையாளம் காணப்படக் கூடியதாகியது. ஐந்தாம் தலைமுறை, பெயர்களிலும், பெயர்களை உச்சரிக்கும் விதங்களிலும் மாற்றமடைந்தது. அதற்கு அடுத்தத் தலைமுறை மத வழிபாட்டு முறைகளில் மட்டுமே தமிழன் என்று அடையாளம் காணப்படக்கூடிய அளவிற்கு மாறியது.

மொழியின் சிதைவுக்கும், மத மாற்றங்களுக்கும், பல் இன, மொழி ஆதிக்கங்களுக்கும் பிறகு தமிழர்கள் ஒரளவிற்கேனும் அடையாளம் காணப்படக் கூடியவர்களாகவும், தமது வரலாற்றை மறக்காதவர்களாகவும் இன்று ஆங்காங்கே வாழ்ந்து கொண்டிருப்பது வியப்புக்குரிய ஒன்றாகும்.

மேற்சொன்ன சூழ்நிலையை ஆராயும் போது முதலில் தமிழர்கள் தமது இரண்டாவது தலைமுறையிலேயே வரிவடிவ எழுத்துத் தமிழை இழந்து விட்டார்கள். அதன் காரணமாகவே தொடர்ந்து தமிழினம் தன்னுடைய இயல்புகளை இழந்தது.

அண்மையில் அறிவும், ஆற்றலும் உள்ள தமிழர்கள் தொழில் நெறிஞர்களாக (Professionals) அமெரிக்க ஐக்கிய நாடு, இங்கிலாந்து, கனடா, ஆஸ்திரேலியா மற்றும் பல ஐரோப்பிய நாடுகளில் குடியேறியுள்ளனர். அதுபோலவே, இலங்கையில் ஏற்பட்ட இனக் கலவரத்திற்கு பின் தாய் மண்ணில் வாழ முடியாத தமிழர்கள் கனடாவிலும் ஆஸ்திரேலியாவிலும் பல ஐரோப்பிய நாடுகளிலும் குடியேறியுள்ளனர். மேலும் தொழில் நெறிஞர்களாகவும், தொழிலாளர்களாகவும் வளைகுடா நாடுகளிலும், சென்ற 50 ஆண்டுகளாக குடியேறியுள்ளனர். இவ்வாறு குடியேறிய நாடுகளில் தமிழ்க் கல்வி கற்க முடியவில்லை. அவர்களின் இரண்டாவது தலைமுறையினரும் எழுத்து தமிழை இழந்து வருகின்றனர். தமிழ்நாட்டின் செய்திகளையும், அதன் தொடர்புகளையும், தமிழை படிக்கக்கூடிய ஆற்றலையும் இழந்து வருகிறார்கள். மொழியின் வளத்தையும், தாய்த் தமிழகத்தின் அல்லது தாங்கள் வசிக்கும் நாட்டின் அறிவியல், பொருளாதார, பண்பாட்டுக் கலாச்சார நிகழ்வுகளையும் தமிழ் மொழியின் மூலம் அறிந்து கொள்ள முடியாமல் மாற்று மொழி மூலமாக அறிந்து கொள்ள வேண்டிய கட்டாயமும் அவர்களுக்கு ஏற்பட்டுள்ளது.

மலேசியா, சிங்கப்பூர், தமிழ்நாடு போன்ற தமிழர்கள் பெருமளவில் வாழும் நாடுகளிலும் ஆங்கில மொழியின் தேவையும், தாக்கமும், பொருளாதார காரணங்களுக்காக பயன்படுத்த வேண்டிய சூழ்நிலையும் உள்ளன. இவர்கள் தங்களின் பொருளாதார வாழ்வுக்கு தமிழ் பயன்படாது எனும் முடிவிற்குத் தள்ளப்பட்டார்கள். ஆங்கிலம் மாற்று மொழியானது. **எந்த ஒரு மொழியும் அந்த இனத்தின் பொருளாதார வாழ்விற்குப் பயன்படாவிட்டால் அந்த மொழி காலத்தால் அழியும். மொழி அழிந்தால் இனமும் அழியும்.** மேற்சொன்ன விஷயங்களில் மூன்று கருத்துக்கள் மையப்படுத்தப்பட்டுள்ளன.

1. **கலாச்சாரப் பண்பாட்டுத் தொடர்புகளை இழந்தது.**
2. **தமிழ் மொழியின் வரிவடிவத்தை இழந்தது, தமிழைப் படிக்கும் ஆற்றலை இழந்தது.**
3. **தமிழர்களின் பொருளாதார உயர்வுக்கு தமிழ்மொழி பயன்படாத சூழ்நிலையும், தமிழர்கள் பொருளாதார வல்லமை பெறாதவர்களாக இருப்பதும்.**

இந்நிலைகளுக்கு நாம் தீர்வு கண்டாக வேண்டும். முதலில் இணையத்தின் வழியாக உலக தமிழர்களிடையே கலாச்சார பண்பாட்டு தொடர்புகளை எவ்வாறு ஏற்படுத்தலாம் என்பதை பார்ப்போம்.

## இணைய ஊடகம்

இன்றைய தகவல் யுகத்தின் அதி நவீன எளிமையான, செலவு குறைந்த தகவல் தொடர்பு தொழில்நுட்பமான, இணையம் உலகெங்கும் உள்ளவர்களை தொடர்பு கொள்ள வழி செய்யும் அதே வேளையில், அனைத்து வகையான ஊடகங்களை (Media) உள்ளடக்கிய பல் ஊடகமாகவும் விளங்குகிறது.

வானொலியும், தொலைக்காட்சியும், அச்சில் வரும் அனைத்து வகையான பத்திரிகைகளும், உலகின் ஒரு குறிப்பிட்ட பகுதியை மட்டும் சென்றடைய கூடியதாக உள்ளது. செய்திகளும் தகவல்களும் பகுதி அல்லது நாடு சார்ந்த அளவிலேயே உள்ளன. உலகத் தமிழர்கள் அனைவரின் தேவையையும் பூர்த்தி செய்யக் கூடியதாக இல்லை. இன்றைய ஊடகங்கள் ஒரு வழி செய்தி தொடர்பாகவே உள்ளது. இருவழி கருத்து பரிமாற்றத்திற்கு வாய்ப்பு இல்லாமல் இருப்பது மற்றொரு குறைபாடாகும். உலக தமிழர்களை இணைக்கக் கூடிய வல்லமை பெற்றதாகவும் இரு வழிச் செய்தி பரிமாற்றம் கொண்டதாகவும் அமைந்திருப்பது இணையத்தின் சிறப்பாகும்.

1. **இணையத்தின் வழி கொண்டு தமிழ், தமிழர், தமிழர் வாழும் நாடுகள் தொடர்பான அனைத்து செய்திகளும் தகவல்களும் தொகுக்கப்பட வேண்டும்.**

    தமிழர்களின் தாயகங்கள் பற்றிய வரலாறு, தமிழர்கள் பல்வேறு நாடுகளில் சென்று குடியேறிய வரலாறு. வாழ்க்கைச் சூழ்நிலை, கல்வி, பொருளாதார நிலைமைகள் தமிழர்களின் ஊடகங்கள், தமிழ் மொழியின் நிலை, வெளிவரும் நூல்கள், கலை, கலாச்சார, பண்பாட்டு, அரசியல், கூறுகள். தமிழிசை, நாட்டியம், ஓவியம், சிற்பம், கட்டடக்கலை பற்றிய தகவல்கள். தத்தம் துறையில் சிகரம் தொட்ட தமிழர்கள், தமிழ் தமிழர் அமைப்புகள், அதன் செயற்பாடுகள், வழிபாடுகள் - தொடர்பான செய்திகள், தொழில் வாய்ப்புகள், நாடுகளின் ஒவ்வொரு பகுதியும் நகரங்களும் பெற்றுள்ள தொழில், வணிக, சுற்றுலா வாய்ப்புகள் போன்ற எண்ணற்ற தலைப்புகளில் தகவல்கள் தொகுக்கப்பட்டு இணையத்தின் வழி வழங்கப்பட வேண்டும். தொகுக்கப்படும் அனைத்து செய்திகளும் இணையம் வழி உலகத் தமிழர்களைச் சென்றடையும்.

இவ்வகையில் தொகுக்கப்படும் அனைத்துச் செய்திகளும் பெருமளவில் மாறுதல் அடையாத (More of Static Data) தகவல்களாகும். **இந்த வகையில் செயல்பட்டு வரும் எமது இணையப் பரப்பு www.intamm.com/infopool.htm**

2. **உலகத் தமிழர்களுக்காக ஒரு BBC, CNN, ராய்ட்டர் போன்ற செய்தி நிறுவனம் உருவாக்கப்பட வேண்டும்.**

 மலேசியத் தமிழர்கள் தமிழ்நாட்டுச் செய்திகளைப் படிக்கின்றனர். ஆனால் மலேசியச் செய்திகளை சிறிதளவே தமிழ்நாட்டு தமிழர்கள் படிக்கின்றனர். அதிலும் தமிழர்கள் தொடர்பான செய்திகள் குறைவே. அதைப்போலவே மொரிஷியஸ், பிஜி தீவு தமிழர்களின் நிலையைப் பற்றித் தமிழ்நாட்டுத் தமிழர்கள் அறிய வாய்ப்பில்லாமல் உள்ளது. மொரிஷியஸ் தமிழர்களுக்கு மிக மிக குறைந்த அளவே தமிழ்நாட்டு தகவல்கள் சென்றடைகின்றன. இந்நிலை மாறவேண்டும். அனைத்து நாட்டுத் தமிழர்களுக்கும், தமிழ், தமிழர் தொடர்பான செய்திகள் உலக அரங்கில் பரவலாக்கப்பட வேண்டும். எனவே இணையத்தின் மூலம் ஒவ்வொரு நாட்டிலும் வாழும் தமிழர்கள் சிலர் இணைக்கப்பட்டு அவர்கள் மூலம் அந்தந்த நாட்டின் தமிழர்களுக்கு ஆர்வம் உள்ள அரசியல், பண்பாட்டு, காலாச்சார, பொருளாதார செய்திகள் தொகுக்கப்பட்டு இணையத்தில் மையப்படுத்தப்பட வேண்டும். இந்த அமைப்பு தமிழர்களுக்கான ஒரு செய்தி நிறுவனமாக இயங்கும். இதன் வழி அனைத்து ஊடகங்களும் (வானொலி, தொலைக்காட்சி, அச்சில் வெளிவரும் பத்திரிக்கைகள்) பயன்பெற முடியும். இவ்வகையில் தொகுக்கப்படும் அனைத்து செய்திகளும் அன்றாடம் மாறக்கூடிய (Dynamic Data) தகவல்களாகும். **இந்த நோக்கில் செயல்படும் எமது இணைய முகவரி www.intamm.com/magazines.htm**

3. **இணையத்தின் மூலம் வானொலி, தொலைக்காட்சி, பத்திரிகைகள் நடத்தப்பட வேண்டும்.**

 இணையத்தின் வழி நடத்தப்படும் வானொலி, தொலைக்காட்சி பத்திரிகைகள் உள்ளடக்கத்திலும் பரவலிலும் உலகம் தழுவியதாக அமையும். இன்றைய வானொலி, தொலைக்காட்சி, பத்திரிகைகள் போன்ற ஊடகங்களிலும் உள்ள குறைபாடுகள் இணையத்தின் மூலம் நடைபெறும் போது இல்லாமல் செய்ய முடியும். மேலும் செழுமையாக்கவும் முடியும். இதற்காக நாங்கள் பல இணைய இதழ் முகவரிகளை உருவாக்கியுள்ளோம், சில

 **www.intamm.com/winnosai/ionline.htm**
 **www.intamm.com/ta-comp/index.htm**
 **www.intamm.com/valar/contents.htm**

நாடகங்கள், திரைப்படங்கள், செய்திப்படங்கள் மற்றும் பல்வேறு வகையான நிகழ்ச்சிகளையும் இணையத்தின் வழி கேட்கவோ, பார்க்கவோ வாய்ப்புள்ள நிலைகள் உருவானால் கலாச்சார, பண்பாட்டு நிகழ்வுகளில் காலத்தாலும், தூரத்தாலும் ஏற்படக்கூடிய மாறுதல்கள் மட்டுப்படுத்தப்படும். **வலுவான உலகத்தமிழர் கலாச்சார பண்பாட்டு அடித்தளம் அமையும்.** பார்வைகளும், தேவைகளும் ஒரு முனை நோக்கி அமையும். இழந்த கலாச்சார பண்பாட்டு தொடர்புகளை மேற்சொன்ன மூன்று செயற்பாடுகளின் மூலம் நம்மால் நிச்சயமாக இணையத்தின் துணை கொண்டு அமைத்து விட முடியும்.

### தமிழ் படிக்கும் ஆற்றல்

இரண்டாவது சிக்கலாகிய, தமிழ் எழுத்தை படிக்கும் ஆற்றலை நமது அடுத்த தலைமுறை இழப்பதனால் அவர்களுடன் நெருக்கத்தையும் தொடர்புகளையும் நாம் இழக்கின்றோம்.

கணிணி வழியாக பல்-ஊடக தொழில் நுட்பத்தின் துணை கொண்டு (Multimedia Technology) மென் பொருட்கள் உருவாக்கப்பட வேண்டும். பல மென் பொருட்களும் உருவாகியுள்ளன. அவ்வகையில் எங்களால் உருவாக்கப்பட்டுள்ள குறுந்தகடு **"செந்தமிழ்"** என்னும் பெயர் உடையதாகும். செந்தமிழ் மூலம் ஆசிரியர் சென்று பயிற்றுவிக்க இயலாத சூழ்நிலையில் உள்ளவர்களுக்கு கணிணியின் மென் பொருளே ஆசிரியராக இருந்து ஒலி, ஒளி, படம் எழுத்து போன்ற அனைத்து ஊடகங்களையும் பயன்படுத்தி மிகவும் எளிமையாகவும்; விரைவாகவும் பயிற்றுவிக்க முடியும். குறுந்தகடு (CD ROM) வழியாகவோ அல்லது இணையம் வழியாகவோ உலகின் எந்த மூலையிலும் வாழும் தமிழ் மாணவர்கள் தமிழ் பேச்சு, எழுத்து, இலக்கணத் தமிழை முறையாகக் கற்று விட முடியும். இணையத்தின் வழியாகவோ அல்லது பத்திரிகை, வானொலி, தொலைக்காட்சி நிகழ்ச்சிகளைப் படிக்கும் போதும், கேட்கும் போதும், பார்க்கும் போதும் அவர்களுடைய மொழித்திறன், சொற்திறன் செழுமை பெறும். தமிழர்களுடைய மொழி, பண்பாட்டு கலாச்சாரம் ஆகியவை உலக அளவில் பேணப்படும்.

## உலகத் தமிழர்கள் பொருளாதார வல்லமை பெற

மூன்றாவதும் மிக முக்கியமான சிக்கலுமாகிய தமிழர்களின் பொருளாதார உயர்வுக்கு தமிழ் மொழி பயன்படவேண்டும். தமிழர்கள் பொருளாதார வல்லமை பெறவேண்டும் என்பதற்கு எப்படி இணையத்தை பயன்படுத்தலாம். எந்த ஒரு பிரச்சனைக்கும் தீர்வு காண்பதற்கு முன், நம்முடைய பலம் என்ன, பலவீனம் என்ன, என்பது ஆராயப்பட வேண்டும். அந்த வகையில்

## தமிழர்களின் பலம்

1. தமிழர்களின் பலம், ஆறு கண்டங்களிலும், அறுபத்தைந்து நாடுகளுக்கு மேற்பட்டு பரவி வாழும் ஒன்பது கோடித் தமிழர்கள் ஆகும்.

2. கல்வி கற்ற தொழில் நெறிஞர் சமூகமாக தமிழர்கள் விளங்குகின்றனர்.

   நூற்றாண்டுகளாக தமிழர்கள் இழந்த கல்வியை, தமிழ் நாட்டில் தந்தை பெரியார், காமராசர் போன்றவர்களின் அயராத உழைப்பின் காரணமாகவும், மற்ற நாடுகளில் உள்ள தமிழர்கள் அந்தந்த நாட்டின் பொருளாதார வளர்ச்சி, அரசியல் சூழ்நிலைகளின் காரணமாகவும் பெற்று, கல்வி கற்றவர்களாக இந்த நூற்றாண்டில் மாறியுள்ளனர். மொத்தத்தில் உலக அரங்கில் கல்வி கற்ற தொழில்நெறிஞர் (Professional Community) இனமாக தமிழர்கள் விளங்குகிறார்கள்.

இவை இரண்டும் மிக சிறப்பான பலங்களாகும்.

## தமிழர்களின் பலவீனம்

1. உலகம் பரவிய தமிழர்கள் தங்களுக்குள், அறிமுகமும், தொடர்பும், பொருளாதார உறவுகளும் இல்லாமல் இருப்பது.

2. உலகம் பரவிய தமிழர்களின் அறிவும், ஆற்றலும், உழைப்பும் தமிழ் இனத்தின் வளர்ச்சிக்குப் பயன்படாமல் இருப்பது.

நாம் இன்றையத் தமிழர்களின் வாழ்வையும், வரலாற்றையும், பலத்தையும், பலவீனத்தையும் பார்த்தோம். பலவீனங்கள், பலமாக மாற்றப்பட வேண்டும். எண்ணிக்கை பலம் எண்ண பலமாக மாற்றப்பட வேண்டும். எண்ண பலம், பொருள் வளமாக மாற்றப்பட வேண்டும். நமது விடியலுக்கான தீர்வுகளையும் வரலாறு தந்திருக்கும் பாடங்களில் இருந்தே பெற்றுக் கொள்ளலாம்.

## தொழிற்புரட்சியிலிருந்து தகவல் யுகத்திற்கு

இங்கிலாந்தில் பொருள் உற்பத்தி முறையில் ஏற்பட்ட மாறுதலின் காரணமாக உருவானது தொழிற்புரட்சி. தொழிற் புரட்சியினால் ஏற்ப்பட்ட உற்பத்தி பெருக்கதால் சந்தை விரிவடைந்தது, புதிய சந்தைகளை தேடி புறப்பட்ட ஆங்கிலேய இனம் தொழில் சார்ந்த வணிக இனமாக வளர்ச்சி பெற்றது. உலகில் அந்தக் காலக் கட்டத்தில் ஏற்பட்ட அந்த மாறுதலை உள்வாங்கிக் கொள்ளாத காரணத்தினால் தமிழ்ச் சமுதாயம் கூலிகளாகப் உலகம் பரவிய நிலை ஏற்பட்டது. எனவே **தமிழ்ச் சமுதாயத்தின் ஒரு 20% விழுக்காடு மக்களேனும் வணிகர்களாக, தொழில் அதிபர்களாக மாற்றப்பட வேண்டும்.**

இன்றைய உலகில் பொருளாதார வளம்மிக்க மனிதர்களே மதிக்கப்படுகின்றனர். எக்காலத்திலும் பிச்சைக்காரர்கள் மதிக்கப்படுவதில்லை. உலக அரங்கில் தமிழினம் உரிய மரியாதையைப் பெற பொருளாதார வல்லமை பெற்ற இனமாக மாற வேண்டும்.

இன்றைய உலகம், ''உலக மயமாக்கல்'' (GLOBALIZATION) எனும் பொருளாதாரச் சிந்தனையோடு இயங்குகிறது. உலகம் தழுவிய வர்த்தகம், கண்டம் கடந்த பன்னாட்டு நிறுவனங்கள் இயல்பாகி விட்டன. நாம் வாழ்ந்து கொண்டிருப்பது தகவல் யுகம். அனைத்துத் தகவல்களும் திரட்டப்பட்டு, தொகுக்கப்பட்டு, அலசப்பட்டு, அவற்றில் இருந்து புதிய சிந்தனைகளும், முயற்சிகளும், தொடர்புகளும் உருவாகி வரும் காலம் இது. தகவல் யுகத்தின் தேவையின் காரணமாக உருவாகியுள்ள இணையம் ஒரு மாபெரும் சக்தியாகும். நாம் இதுவரை குறிப்பிட்ட தமிழ்ச் சமுதாயத்தின் அனைத்துச் சிக்கல்களுக்கும் இணையத்தின் வழி தீர்வுகாண முடியும்.

## அறிவும் ஆற்றலும்

உலகத் தமிழர்கள் பொருளாதாரத்தில் வல்லமை பெற்றவர்களாகத் திகழ வேண்டும். ஒரு சமுதாயம் பொருளாதார வெற்றிக்காண வேண்டுமானால் அறிவும், ஆற்றலும், பணமும், பொருளும் ஒன்றாக இணைந்து புதிய தொழில் நுட்பங்களைக் கைக் கொண்டு, புதிய தொழில் முயற்சிகளில் ஈடுபட வேண்டும். வணிகச் சந்தை விரிவுப்படுத்தப்பட வேண்டும். ஆனால் தமிழ்ச் சமுதாயத்தின் அறிவும்,

ஆற்றலும், உழைப்பும் மற்ற இனத்தவரின் உயர்வுக்கு கூலிக்கு விற்கப்பட்டுக் கொண்டிருக்கிறது. படிக்காத கூலிகளாக இருந்த தமிழினம் இன்று படித்தக் கூலிகளாக மாறிவருகிறது.

தமிழர்களுடைய அறிவு, ஆற்றல், உழைப்பை பயன்படுத்தக் கூடிய, வளர்க்கக் கூடிய ஆற்றல், இன்றைய தமிழ் இனத்திற்குப் போதுமானதாக இல்லை. **ஒரு இனத்தின் அறிவும், ஆற்றலும், உழைப்பும் அந்த இனத்தின் உயர்வுக்கு பயன்படாவிட்டால் அந்த இனம் உயர்வு காண்பது எங்ஙனம்?**

தமிழினம் உயர்வு பெற சிந்தனைப் போக்கில் மாறுதல் வேண்டும். புதிய தொடர்புகள், புதிய தொழில் நுட்பங்கள், புதிய தொழில் முயற்சிகள், புதிய சந்தைகள் கண்டறிந்து உருவாக்கப்பட வேண்டும். தமிழ்ச் சமுயதாயத்தின் அறிவும், ஆற்றலும் தமிழ் இனத்தின் உயர்வுக்கு மிகப் பெரிய துணையாக இருக்கும்.

## தொழில் நெறிஞர் பேரேடு (Professionals Directory)

உலகெங்கும் வாழும், அறிவும், ஆற்றலும் உடைய தொழில் நெறிஞர்கள் புதிய தொழில் நுட்பங்களை மற்ற நாடுகளில் வாழும் தமிழர்களுக்கு அறிமுகப்படுத்தக் கூடிய வாய்ப்பும் வசதியும் பெற்றிருக்கிறார்கள்.

எனவே இத்தகைய தொழில் நெறிஞர்கள் அடையாளம் காணப்பட்டு அவர்கள் பெயர், முகவரி, சிறப்பு பெற்ற துறை, சாதனைகள் போன்ற செய்திகள் இணையத்தின் வழியாக பேரேடாக (Directory) தொகுக்கப்பட்டு உலகத் தமிழர்களுக்கு அறிமுகப்படுத்தப்படும். இதன் மூலம் புதிய தொடர்புகள் ஏற்பட வழி வகை செய்யப்படும். புதிய தொழில் முயற்சிகளுக்கான தொழில் நுட்பங்கள் மட்டுமன்றி துறை சார்ந்து அல்லது துறை கடந்து தகவல் பரிமாற்றம் தொழில் நெறிஞர்களுக்கிடையே ஏற்படவும் இப்பகுதி வழி செய்யும். தகவல் பரிமாற்றம் தொழில் நெறிஞர்களை மேலும் சிறப்புப் பெற்றவர்களாக மாற்றும். அதன் வாயிலாகப் பல்வேறு பயன்கள் நிகழும். நாங்கள் உருவாக்கியுள்ள **இணைய முகவரி www.intamm.com/prof_link.htm**

## வணிகப் பேரேடு (Business Directory)

சமுதாயத்தின் ஒரு பகுதியாக விளங்கும் தொழில் நெறிஞர்கள் போன்று, மற்றொரு பகுதியான பணமும் பொருளும் உடைய வணிகர்கள், தொழில் அதிபர்கள். தொழில் நிறுவனங்கள் அடையாளம் காணப்படவேண்டும். அவர்கள் வாங்கும், விற்கும், ஏற்றுமதி, இறக்குமதி செய்யும் பொருட்கள், உற்பத்தி செய்யும் பொருட்கள் அவை பற்றிய விவரங்கள் ஆகியவை தொகுக்கப்பட்டு வணிகப் பேரேடு இணையத்தில் உருவாக்கப்படவேண்டும். இது ஒரு பொருள் சார்ந்த பேரேடாக (Product Directory) அமையும். இதன் மூலம் ஒரு குறிப்பிட்ட பொருளை உலக அரங்கில் கையாளும் நிறுவனம், அதன் தொடர்பு முகவரி உட்பட அனைத்துச் செய்திகளையும் தேடுதல் முறையில் (Searchable) நாம் பெறக் கூடிய சூழ்நிலை உருவாக வேண்டும். வணிகப் பேரேடாக அமைந்துள்ள இணைய முகவரி **www.intamm.com/busi_link.htm**

## தொழில் நெறிஞர் பேரேடும் வணிகப் பேரேடும்

தொழில் நெறிஞர் பேரேட்டின் மூலம் தொழிலதிபர்களும், வணிகர்களும் தொழில் முனைவோர்களும் ஒருவருக்கொருவர் தொடர்பு கொண்டு புதிய தொழில் நுட்பங்களையும், தொழில் அறிவையும் பெற்றுக் கொள்ள முடியும்.

வணிகப் பேரேட்டின் மூலம் ஒரு தொழில் நெறிஞர் ஒரு குறிப்பிட்ட தொழில் நிறுவனத்தை அணுகி தங்களின் ஆற்றல் மற்றும் அனுபவத்தின் துணைக் கொண்டு வழிகாட்டலாம். தொழில் நெறிஞர்களும் பொருளீட்டலாம். இதன் மூலம் தொழில்கள் சிறப்பாக நடைபெறும்.

தொழில் முனைவோருக்கு தகுந்த வழிகாட்டுவதோடு, தகுதிவாய்ந்தவர்களுக்கு முதலீட்டு வாய்ப்புகளை (Investment Opportunities) அறிமுகப்படுத்தக்கூடிய வாய்ப்பும் ஏற்படும்.

தொழில் நெறிஞர் பேரேடு, வணிகப் பேரேடு இவற்றின் துணையால் புதிய தொழில் நுட்பங்கள் கொண்டு தொழில்கள் செழுமைபெறும், புதிய தொழில் முயற்சிகள் ஏற்படும். பல புதிய வேலை வாய்ப்புகள் உருவாக்கப்படும். அதுமட்டுமில்லாமல் இந்தப் பகுதிகளின் மூலம் தகுதிவாய்ந்த தமிழர்கள் வெளிநாடுகளிலும் வேலை வாய்ப்பு பெறக்கூடிய சூழ்நிலையும் ஏற்படும். அவ்வகையிலும் தமிழர்கள் பொருளாதார வல்லமை பெறுவார்கள்.

## தமிழருக்கான உலகச் சந்தை

புதிய தொழில்கள் மலர்வதன் தொடர்ச்சியாக உற்பத்தி பெருகும். தமிழரின் சந்தை விரிவடையும். தேவை ஏற்படும். அச் சூழ்நிலையில் பல நாடுகளிலும் வாழும் தமிழர்கள் தம் நாடுகளுக்கு தேவையான பொருட்களை வாங்கி விற்க கூடிய (Trading) வாய்ப்பு உருவாகும். பல நாடுகளில் தமிழரின் பொருட்கள் விற்பனைக்கு செல்லும். அந்த அந்த நாடுகளில் உற்பத்தியாகும் ஏனைய பொருள்களும் உலகச் சந்தையில் இடம்பெறுவது தமிழர்கள் வழி நிகழும்.

இது தொடர்பாக நிகழக்கூடிய அனைத்து, தகவல் தொடர்புகளுக்கும் செலவு குறைந்த, விரைவான, எளிமையான தகவல் தொடர்பாகிய இணையம் பெருமளவு பயன்படும்.

## வணிக மொழி தமிழ்

இன்றளவில் உலகில் பிற தொழில், வணிகம் சார்ந்த இனங்கள் தங்கள் இனத்தை சேர்ந்தவர்களோடு தொடர்பு கொண்டு வணிகம் செய்யும் போது தங்களின் தாய் மொழியிலேயே தொடர்பு கொள்கிறார்கள், உரையாடுகிறார்கள். அந்த வகையில் தமிழர்களுக்கிடையே வணிகம் நடைபெறும் போது தமிழ் மொழி வணிக மொழியாக மாறும்.

பொருளாதார வல்லமை பெற்றவர்களின் மொழி, பிறிதொரு மொழியை தாய் மொழியாக பெற்றவர்களால் பொருளாதார காரணங்களுக்காகக் கற்கப்படுகிறது. உதாரணம் ஆங்கிலம், ஜெர்மன், சப்பானியம், பிரெஞ்சு போன்றவை. தமிழர்கள் **World is our resources. World is our Market** என்ற சொற்றொடருக்கு ஏற்ப உலகம் தழுவிய வணிகத்தில் ஈடுபட்டு பொருளாதார வல்லமை பெற்றவர்களாக மாறும் போது இன்று நாம் கற்கத் தயங்கும் தமிழ் மொழி உலகத்தவராலும் கற்கப்படும். இன்று மொழி வளத்திற்காகவும், ஆன்மீக, வாழ்வியல் கருத்துக்களுக்காகவும் பயிலப்படும் தமிழ் மொழி உலக மாந்தர்களால் தமிழர்களோடு ஏற்படக்கூடிய வணிக தொடர்புக்காகவும் கற்கப்படும் நிலை வரும்.

## தமிழர்களுக்காக இணையம்

இக்கட்டுரையில் குறிப்பிடப்பட்டுள்ள பல்வேறு இணையப் பரப்புகள் மற்றும் முகவரிகளின் துணைகொண்டு அரிய தகவல்கள் பெறப்பட்டமை, புதிய தொழில் முயற்சிகள், பொருள்கள் சந்தைப்படுத்தப்பட்ட நிகழ்வுகள், தொழில்நுட்ப தகவல் பரிமாற்றங்கள், புதிய அறிமுகங்கள் கடந்த ஓராண்டுக்குள் உலக தமிழர்களிடையே நிகழ்ந்துள்ளன என்பதை மகிழ்ச்சியுடன் குறிப்பிட விரும்புகிறேன். எனவே உறுதியாக இணையத்தின் மூலம் உலகத் தமிழர்களின் கலாச்சார பண்பாட்டு தளம் வலுவாக அமைக்கப்படும். தமிழர்களிடையே நெருக்கமான தொடர்புகள் ஏற்படும் தமிழ் மொழி வணிக மொழியாகும், தமிழர்கள் அறிவு சார்ந்த துறைகளில் வெற்றி பெறுவார்கள், தமிழர்களின் வணிகம் உலகம் தழுவியதாக அமையும் தமிழர்கள் பொருளாதார வல்லமை பெறுவார்கள். தமிழர்கள் உலகில் உரிய மரியாதையை பெறுவார்கள்.

**தொன்ம வேர்களாகிய தமிழர்களை மின் வேராகிய இணையத்தின் மூலம் இணைக்கும் போது இவ்வுலகில் தமிழர் எழுச்சி ஏற்படும்.**

ISBN 81 - 7023 - 903 - 6

Allied Publishers Ltd.